SANTA ANA RIVER USE-ATTAINABILITY ANALYSIS

VOLUME 6:

REVIEW OF CHRONIC WHOLE EFFLUENT TOXICITY TESTING REQUIREMENTS

Principal Researchers

Chadwick & Associates, Inc.
Littleton, Colorado

and

Regulatory Management, Inc.
Colorado Springs, Colorado

Prepared For:

Santa Ana Watershed Project Authority
Riverside, California

Under Contract with:

Risk Sciences
Colorado Springs, Colorado

September, 1992

TABLE OF CONTENTS

		Page
Section 1:	Introduction	1
1.1	Background	1
1.2	Key Public Policy Questions	2
1.3	Purpose of Review	3
Section 2:	Chronic Test's Ability to Measure the Presence or Absence of Marginal Toxicity	5
2.1	Experimental Design Constraints	5
2.2	Implications of Biological Variability	7
2.2.1	Sub-lethal Endpoints - Reproduction and Growth	10
2.2.2	Mortality Endpoint - Survival	14
2.2.3	Biological Variability - Implications to Test	16
2.2.4	Coefficient of Variation for Control Organisms	19
2.3	Impact of Variability on Test Precision	22
2.3.1	Number of Tests for Desired Precision	23
2.3.2	Compensation for Test Variability	26
2.4	Variability is Inherent to Test	28
2.4.1	Variability not Induced by Laboratory Performance	29
2.4.2	EPA Recognizes the Presence of Variability	30
Section 3:	Chronic Test's Ability to Reliably Predict Impairment to Aquatic Ecosystems	33
3.1	Ability of Chronic WET tests to Predict Toxicity	33
3.2	Effluent Samples Cannot Represent In-Stream Conditions	33
3.3	Test Conditions Cannot Represent In-Stream Conditions	35
3.3.1	Culturing Procedures	35
3.3.2	Dissolved Oxygen Compensation	36
3.3.3	Physical Conditions	37
3.3.4	Analytical Procedures	37
3.4	Test Organisms Are Not Representative Of Santa Ana River Fauna	39
3.4.1	Fathead Minnows and *Ceriodaphnia* Are Not Native	39
3.4.2	Species Give Conflicting Results	40
3.5	UAA Provides A Case Study	43
3.6	EPA Field Validation Studies	44

TABLE OF CONTENTS - continued

		Page
Section 4:	Implications for NPDES Compliance	48
4.1	False Results Reduce Test Utility	48
4.1.1	Implications of False Negative Results	48
4.1.2	Implications of False Positive Results	49
4.3	Inability to Interpret or Complete TIE/TREs	51
4.4	Conflict With Other Indicators of Water Quality	52
4.5	Unenforceable Effluent Limitations	53
4.6	Continuous Liability	54
Section 5:	Conclusions and Recommendations	56
5.1	Conclusions	56
5.2	Recommendations	57
5.3	Summary	59
Literature Cited		61

SECTION 1: INTRODUCTION

1.1 BACKGROUND

California's Inland Surface Waters Plan (ISWP) established water quality objectives for acute and chronic toxicity. The plan calls for regular toxicity testing by exposing fish, invertebrates, and algae to municipal effluent under controlled laboratory conditions. Toxicity is indicated when effluent-exposed organisms fail to survive, grow, or reproduce at a level which is statistically equivalent to similar organisms which were not exposed to effluent.

Under California state regulations, "consistent" failure of chronic whole effluent toxicity (WET) tests triggers a requirement to conduct a Toxicity Identification Evaluation (TIE) and, if necessary, a Toxicity Reduction Evaluation (TRE). Failure to execute a TIE and/or TRE subjects the discharger to potential fines, imprisonment, and permit revocation.

Recently, U.S. Environmental Protection Agency (EPA) objected to new NPDES permits issued to implement California's Inland Surface Waters Plan. The permits were issued to six POTW's discharging to the Santa Ana River System: San Bernardino, Colton, Rialto, Riverside, and the Chino Basin Water Reclamation District.

EPA opposed the State of California-issued permits because the permits do not make toxicity test failures "per se" permit violations subject to full enforcement under the law. EPA maintained that the ISWP provisions, which allow Publicly Owned Treatment Works (POTWs) to avoid civil and criminal penalties by conducting a TIE/TRE, do not adequately implement federal Clean Water Act requirements.

EPA threatened to veto any permit which fails to make exceedence of the ISWP limitation on chronic toxicity (1TUc) an independently enforceable permit violation. EPA also indicated that they would issue federal NPDES permits if necessary.

The affected dischargers in the Santa Ana River basin disagreed with EPA's approach on the basis that the sub-lethal endpoints (reproduction and growth) of the chronic WET test are not sufficiently robust to determine permit compliance. They also claimed that inflexible enforcement would tend to discourage Toxicity Identification and Reduction Evaluations.

During the conduct of a use-attainability analysis and development of site-specific water quality objectives for the Santa Ana River, there were many occasions where various indicators (chemical exceedences, biomonitoring failures, and instream impairment assessments) gave conflicting signals regarding the presence or absence of toxicity and impairment of biological communities in-stream.

In-depth investigation of the discrepancies identified considerable, heretofore unreported, variability in the sub-lethal endpoints of the controls from chronic WET tests. Such background variability must be compensated for in the test procedures and interpretive statistics before valid conclusions about the presence or absence of toxicity can be derived.

1.2 KEY PUBLIC POLICY QUESTIONS

The ISWP toxicity objectives disallow POTWs from discharging effluent with any apparent adverse impact on aquatic life. It is a "zero-tolerance" standard; any indication of toxicity is deemed unacceptable under the law.

When the California Water Resources Control Board adopted the ambitious 1TUc limit, they did so to protect the ISWP narrative water quality criteria which prohibits "toxics in toxic amounts" from being discharged to waters of the state. They also established procedural criteria which emphasize cleanup and abatement (TIE/TRE) over more traditional penalties.

The State of California relied on EPA's claim that the sub-lethal endpoints, which are specified in the recommended test protocols for chronic WET testing, were valid indicators of chronic toxicity and in-stream impairment. Even if the chronic WET test was less than 100% reliable, the requirement that toxicity be "consistently" present before triggering any follow-on

TIE/TRE would provide sufficient opportunity to distinguish and dismiss spurious test results. However, when EPA forced the Santa Ana dischargers to accept permits which made every chronic WET test failure a violation, before the presence of toxicity could be confirmed and the cause of failure could be corroborated through TIE, the question of test reliability became more critical.

The strict liability requirements of the Clean Water Act, the zero-tolerance nature of California's 1 TUc objective, and the documented variability of the sub-lethal endpoints of the chronic WET test (see discussion below), make it vital to address these key policy questions: 1) should the sub-lethal endpoints of the chronic WET test, as expressed in current guidance and procedures, be used to determine compliance with NPDES permits? 2) can the sub-lethal endpoints of the chronic WET test **reliably** distinguish the presence or absence of toxicity? and 3) do the sub-lethal endpoints of the chronic WET test **reliably** predict impairment to aquatic ecosystems?

1.3 PURPOSE OF REVIEW

The biomonitoring protocols were originally established as monitoring tools to facilitate the identification and investigation of chronic toxicity. While adequate to perform that task, it is uncertain whether the bioassay tests are sufficiently robust to serve as a pass-fail method for assessing permit compliance (in the same manner as chemical tests).

EPA has not submitted their recommended chronic WET testing protocols to formal scientific peer review nor have standard test methods been adopted into 40 CFR Part 136. Therefore, the purpose of this review and analysis was to:

- Review the scientific literature supporting the use of chronic WET testing. Although not a true "peer review," EPA notes (in the TSD Responsiveness Summary) that the informal review process is an adequate substitute for evaluating the chronic WET test protocols.

- Review available laboratory and field data to determine the power of chronic WET testing to identify toxicity. Particular emphasis was placed on determining the "normal" biological performance of aquatic organisms used in the chronic WET tests in order to establish a reference baseline.

- Recommend adjustments to chronic WET testing protocols to increase their reliability and validity. More robust procedures would significantly reduce the incidence of false toxicity indications and improve the ability of the test to accurately predict aquatic ecosystem impairment.

The analysis is divided into sections. Section 2 focuses on the whether the chronic WET tests can *reliably* distinguish the presence or absence of toxicity, particularly with regards to the sub-lethal endpoints of reproduction and growth. Section 3 investigates whether the chronic WET tests can *reliably* predict impairment to aquatic ecosystems, with emphasis on the evaluation of the sub-lethal endpoints. Section 4 evaluates the implications for NPDES compliance determination. Section 5 summarizes the review and proposes recommendations to improve the utility of chronic WET testing. Supporting literature is cited in Section 6.

SECTION 2: CHRONIC WET TEST'S ABILITY TO MEASURE THE PRESENCE OR ABSENCE OF MARGINAL TOXICITY

2.1 EXPERIMENTAL DESIGN CONSTRAINTS

The purpose of nearly all experiments is to deduce cause/effect relationships based on study results (Weisz 1971). The central organizing principle of robust experimental design is to minimize all extraneous sources of interference and variability. It is axiomatic that all living organisms will tend to vary in the rate at which they reproduce, grow and die. Some level of biological variability is normal and natural in all animals, including those used in scientific experiments.

In order for the biomonitoring test to reliably identify the presence or absence of toxicity, it must be able to distinguish the difference between normal biological variability and changes in the rate of growth, reproduction or mortality caused by contaminated water. Failure to account for natural variations may cause actual toxicity to be obscured by all of the normal fluctuation. Or, the natural variation may be mistaken for toxicity even when the water quality is actually safe.

Chronic whole effluent toxicity test protocols were designed to minimize variability within and between laboratories by standardizing methods through a series of defined protocols (Weber *et al.* 1989, Weber 1991). Many of these protocols were described in the first published versions of the tests (Mount and Norberg 1984, Norberg and Mount 1985) and further developed in a series of "validation" studies conducted by the American Petroleum Institute, Electric Power Research Institute, and others in cooperation with EPA (Cooney *et al.* 1988, 1989, 1992a,b; DeGraeve and Cooney 1987, DeGraeve *et al.* 1989a, 1989b, 1991, 1992). As a result, EPA provides defined protocols for test conditions, including: organism age, species, feeding schedule, amount and type of food, culture conditions, temperature, light, dissolved oxygen, etc. During the test, organisms and chambers are randomized into treatment cells. In addition, in the case of the *Ceriodaphnia dubia* test, it is assumed that the use of identical parthenogenic "clones" would further reduce variance.

Despite the rigorous protocols and apparent uniformity in methods between tests and laboratories, the variability inherent in the use of living organisms has the potential to confound test interpretation and invalidate study conclusions. A number of factors can affect the test results, including genetic variability in the *Ceriodaphnia* and fathead minnow populations, variation in effluent-free laboratory dilution water, feeding and handling techniques, food quality, age of the organisms at test initiation, drying-oven characteristics and the analytical balance (Cooney *et al.* 1992a,b, Cowgill 1987, DeGraeve *et al.* 1991, Patterson *et al.* 1992, Pickering 1988, Snyder *et al.* 1991).

EPA, recognizing that some degree of variability exists in biomonitoring tests, has established control performance criteria to reduce the chance of false negatives (i.e. no measure of toxicity when in fact some exists). Two tiers of criteria must be met for performance of controls, i.e. those organisms exposed only to effluent-free reconstituted laboratory water, before results from a chronic WET test can be accepted (Weber *et al.* 1989, Weber 1991).

The first criterion is for the mortality endpoint: *Ceriodaphnia* and fathead minnow control organisms must exhibit at least 80% survival. The second criterion is for the sub-lethal endpoints of reproduction and growth: *Ceriodaphnia* controls must produce at least 15 young per surviving female and fathead minnow controls must exhibit a mean end-of-test dry weight of 0.25 mg per individual (Weber *et al.* 1989, Norberg-King 1989a). This two-tiered approach is set up to eliminate potential *false negatives* in the tests (e.g. no measure of toxicity when in fact it does exist). This is the standard protection against Type I errors in hypothesis testing (Sokal and Rohlf 1987). However, there is no corresponding system to eliminate potential Type II errors or *false positives* (e.g. a measure of toxicity when in fact none exists).

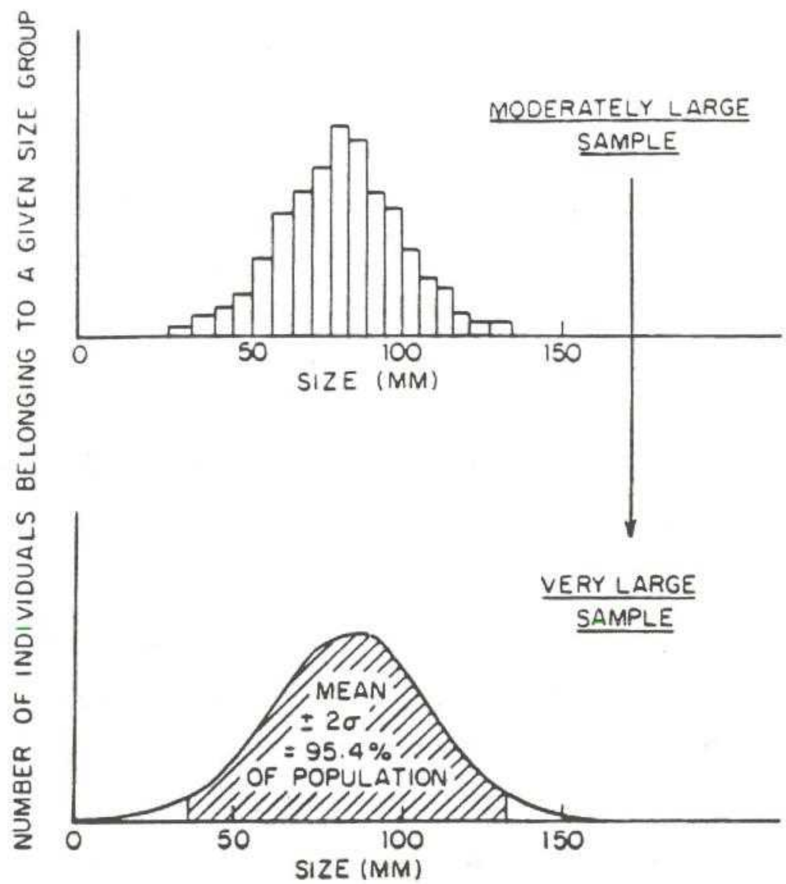
The demonstrated presence of a true dose-response curve (incremental progression of toxicity as percentage effluent increases) in the test dilution series is a third tier of test acceptance criteria recently adopted by EPA (EPA 1991c, Norberg-King 1989b, Mount 1990, Liu 1992), although it is not yet in chronic WET test guidance documents. Using this criterion, if the test dilution series do not follow the expected dose-response curve, the test is considered invalid and should be rejected, "since it would not permit inference of the existence or absence of toxicity" (Norberg-King 1989b, Mount 1990).

In summary, there are cuurently three tiers of criteria and five measures of test acceptance for chronic WET test acceptance/rejection: 1) 80% survival in *Ceriodaphnia* controls, 2) 80% survival in fathead minnow controls, 3) reproduction of control *Ceriodaphnia* $\geq$ 15 per surviving female, 4) growth of fathead minnows resulting in mean end-of-test dry weight $\geq$ 0.25 mg per individual and 5) true dose-response curve test results showing consistent progression of increasing toxicity with increasing concentration of effluent/toxicant. All of these criteria are meant to serve as compensation for the confounding influence of natural biological variability on test interpretation.

2.2 IMPLICATIONS OF BIOLOGICAL VARIABILITY

The variability inherent in biological populations is a strong and, for all practical purposes unmeasured, external source of variance in chronic WET tests. This can be especially true for the "sub-lethal" chronic test endpoints such as *Ceriodaphnia* reproduction (number of offspring per surviving female) and fathead minnow growth (mean end-of-test weight per individual).

It is a basic tenet of population ecology that organisms will exhibit a range of values for a given metric, such as growth or reproduction (Wilson and Bossert 1971, McNaughton and Wolf 1973, Slobodkin 1961). This range of response is often presented graphically in a frequency distribution chart. For many population parameters such as reproduction or growth, the statistical distribution within a given size or age class resembles a bell-shaped curve (Wilson and Bossert 1971), with a few individuals on the low range, a few on the high range, and the majority in the middle (Fig. 1). For a *normal* distribution, approximately 95% of the population will fall within roughly two standard deviations of the mean (i.e. mean $\pm$ t$\cdot$SD, where t = 1.96) (Wilson and Bossert 1971, Snedecor and Cochran 1967, Elliott 1977).



1 FREQUENCY DISTRIBUTIONS of many characteristics of a population yield bell-shaped ("normal") curves. As the number of individuals increases, the frequency curve more closely approaches the ideal form. In a "perfect" normal distribution, 68.3% of all individuals are within one standard deviation of the mean (mean $\pm \sigma$), 95.4% within two standard deviations (mean $\pm 2\sigma$), and 99.7% within three standard deviations (mean $\pm 3\sigma$).

FIGURE 1: Typical frequency distribution curve (from Wilson and Bossert 1971).

It is important to remember that, as with any experiment, the controls are just a sample of the entire population and thus can only represent an estimate of population performance. Ideally, control organisms should perform close to the population mean. If the statistical concept outlined above is carried over to chronic WET tests, it is reasonable to assume that control performance for reproduction, growth or survival should fall within the mean $\pm$ 1.96 SD range for that parameter. In this way, controls would be within the range of normal performance for that metric 95% of the time. It would be unreasonable to require controls to perform at a rate greater than the typical performance for their species. And yet, as will be shown, this is exactly the result of the implementation of the performance criteria set by EPA.

As part the Santa Ana River Use-Attainability Analysis (UAA), a large data base on controls (i.e. organisms exposed only to effluent-free reconstituted laboratory water) has been compiled for both *Ceriodaphnia* and fathead minnows. Chadwick & Associates, Inc. conducted a large number of toxicity tests during the UAA (see Volume 2, Appendix F), providing a considerable data base on control performance.

In addition, we have received results from nine California laboratories conducting chronic WET tests for the Santa Ana River dischargers, including Aqua-Science in Davis, ERC Environmental and Energy Services, Inc., in San Diego, Aqua Terra Technologies in Walnut Creek, Associated Laboratories in Colton, Aquatic Bioassay & Consulting Laboratories, Inc. in Ventura, Ogden Environmental and Energy Services in San Diego, Aquatic Testing Laboratories in Ventura, the City of Riverside's in-house laboratory, as well as Aquatec, Inc. in South Burlington, Vermont. Another major set of data came from the series of chronic toxicity validation studies performed by 13 laboratories under the direction of Battelle Laboratories Columbus Division for the Electric Power Research Institute, American Petroleum Institute, and others in cooperation with EPA (Cooney *et al.* 1988, 1989; DeGraeve *et al.* 1989a, 1989b, 1991, 1992). These 13 laboratories included EPA, private contract, industrial and academic laboratories.

Lastly, control information was used from an inter-laboratory study conducted by San Francisco Bay Regional Water Quality Board using 10 laboratories and additional data reported in the literature (Anderson and Norberg-King 1991, Birge, *et al.* 1989, EPA 1991a). Combined, these studies provide control information from a grand total of 34 laboratories.

The *Ceriodaphnia* data set from the Battelle/EPRI validation studies (DeGraeve *et al.* 1989b) presented only the biomass measure, which is total young produced divided by the total number of organisms originally exposed. These data are not the mean numbers of young produced per *surviving* female needed for the *reproduction* metric. Thus, for the laboratories participating in the Battelle study, the *number of young per surviving female* had to be estimated from the data presented, since the individual replicate information for the complete data set was not included in the document (DeGraeve *et al.* 1989b). Estimates of mean number of young per surviving female were made by dividing the total young by the number of females that survived the test. For example, a data set with 40% survival and a biomass of 22.1 would result in 55.3 mean young per surviving female (221 offspring ÷ 4 surviving females), using the estimation method described above. This method will overestimate the mean young per surviving female in some cases where females had released several young before dying.

When combined with the UAA results, these studies from 34 laboratories from around the country provide over 200 data points on control performance for *Ceriodaphnia dubia* and fathead minnows (*Pimephales promelas*). These data were plotted as frequency distribution charts to allow an estimate of what is "normal" control reproduction, growth and/or survival for these organisms. This, in turn, allowed a determination of where EPA's recommended performance criteria fall, given the normal range of performance by these test organisms.

Because these controls were run with organisms exposed only to reconstituted laboratory water, and not ambient receiving water, there would be no potential for confounding toxic responses to the control water (Norberg-King 1989b). In other words, this analysis provides the opportunity to measure just how many offspring *Ceriodaphnia* produce and how much weight fathead minnows normally gain when water quality is fully supporting.

2.2.1 Sub-lethal Endpoints - Reproduction and Growth

When the results of 214 chronic WET test <u>controls</u> are plotted for *Ceriodaphnia* reproduction, it is apparent that a wide range of values have been reported and a relationship resembling a "bell-curve" normal distribution can be observed (Fig. 2). Given these data, the mean reproduction for *Ceriodaphnia* for a 7-day chronic test control would be 20 young/female with a standard deviation of 9 (Fig. 2). Note that this data set includes the estimated reproduction values based on the data in DeGraeve *et al.* (1989). If the estimated values were not used, the total number of tests would be 184 with a mean of 19 young/female and a standard deviation of 9; virtually identical to the full data set. Using the mean ± 1.96 SD as discussed earlier, the normal range of control reproduction for *Ceriodaphnia* is from 2 to 38 young/female. That is, 95% of the time, *Ceriodaphnia* in reconstituted laboratory water will produce between 2 and 38 offspring per surviving female during a 7-day test, whereas 2.5% of the time they will produce less than 2 offspring and 2.5% of the time more than 38 offspring.

Standard statistical procedures would declare controls that perform within this range of 2-38 offspring as adequate for test acceptance. However, EPA's criteria of 15 young/female falls well above the lower bound of the range of normal performance (expressed as the 95% confidence interval). In fact, approximately 22% of all observations fall below the EPA criteria of ≥ 15 young/surviving female. In addition, application of this criterion of ≥ 15 young/female would invalidate up to 15% of the test controls that still fall within the mean ± 1.96 SD (Fig. 2). In other words, up to 15% of all tests would be rejected for inadequate control performance even though their controls fall within the natural range (95% confidence interval) of *Ceriodaphnia* reproduction. These rejected tests would then have to be re-run at considerable expense for either the discharger or the laboratory.

Since this wide range of variability in *Ceriodaphnia* reproduction was observed in effluent-free, reconstituted laboratory water, it is reasonable to assume that it occurs in the effluent dilution series of a chronic WET test, as well. When comparing number of young produced by *Ceriodaphnia* in effluent dilutions vs. those in reconstituted laboratory water, a chronic WET test failure (measured as reduced reproduction) could well be the result of the natural biological variability of *Ceriodaphnia* reproduction and not actual toxicity (i.e. a false positive).

When the results of 230 chronic WET test controls are plotted for fathead minnow end-of-test weight, it is again apparent that a wide range of values exist (Fig. 2). As with *Ceriodaphnia* reproduction, a relationship resembling a "bell-curve" normal distribution can be observed. The mean end-of-test weight for fathead minnows using these data for 7-day chronic test controls is 0.43 mg per individual with a standard deviation of 0.21 (Fig. 2). Using the

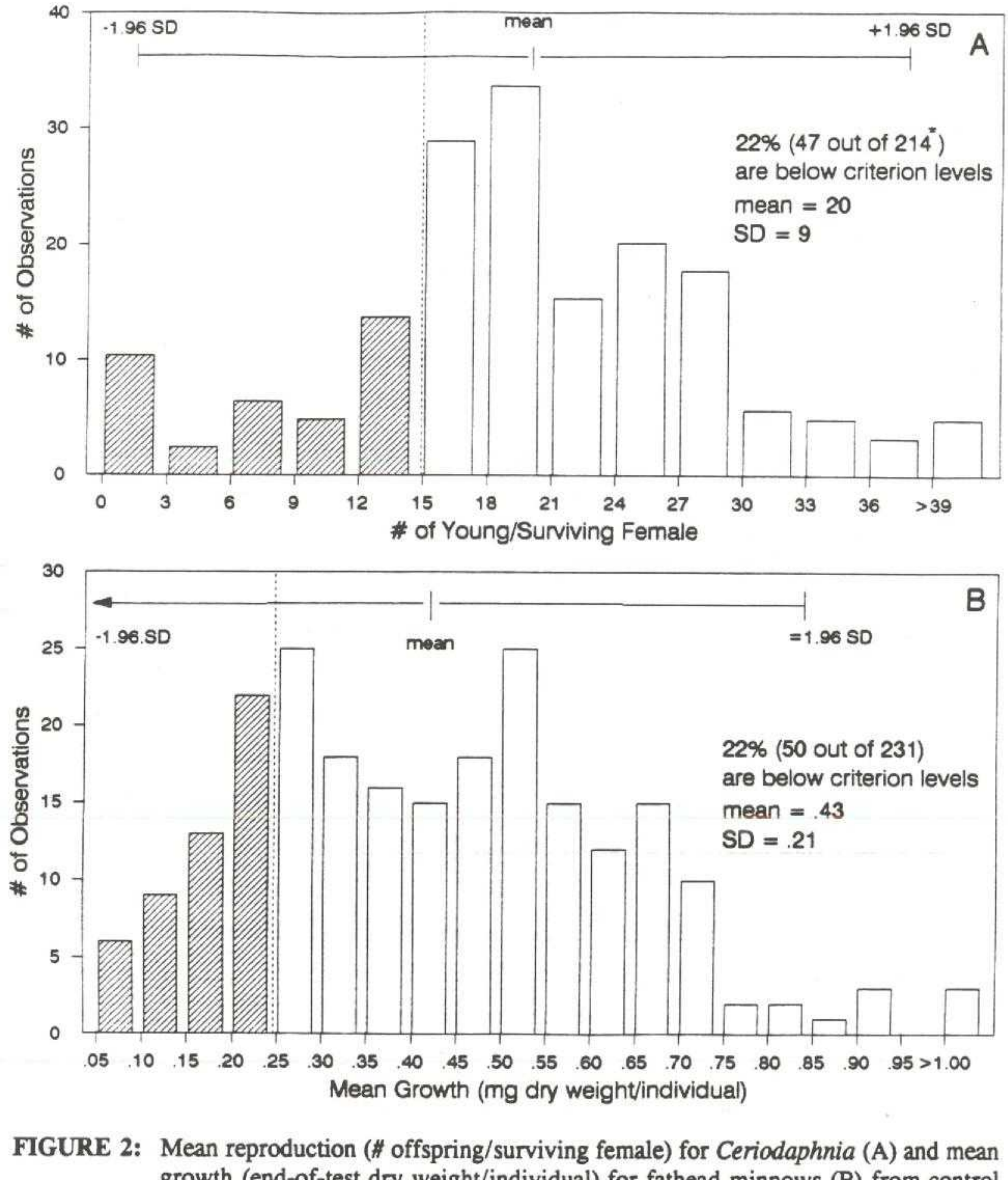


FIGURE 2: Mean reproduction (# offspring/surviving female) for *Ceriodaphnia* (A) and mean growth (end-of-test dry weight/individual) for fathead minnows (B) from control organisms exposed to effluent-free reconstituted laboratory water.
*(includes estimated *Ceriodaphnia* reproduction from DeGraeve, *et al.* 1989b - see text for explanation.)

mean $\pm$ 1.96 SD, the normal range of end-of-test weight for fathead minnow larvae exposed only to reconstituted laboratory water is 0.02 to 0.84 mg/individual.

Using standard statistical procedures, fathead minnow controls that perform within this range would be adequate for test acceptance. However, as was the case with *Ceriodaphnia* reproduction, EPA's fathead minnow control criteria for end-of-test weight of 0.25 mg/individual falls well above the lower bound of the range of normal performance and would invalidate up to 22% of the test controls that still fall within the mean $\pm$ 1.96 SD (Fig. 2). In other words, up to 22% of all tests would be rejected using the EPA control performance criteria of $\geq$ 0.25 mg/individual even though their controls fall within the natural range of fathead minnow end-of-test weight. Any rejected test would have to be re-run.

Because the variability in growth among control organisms is apparently related to normal biological processes, it is likely that organisms selected for exposure to effluent would also exhibit similar variability regardless of the effluent's direct effects. When determining toxicity by comparing end-of-test weights of fathead minnows from effluent dilutions vs. those exposed only to reconstituted laboratory water, this analysis again suggests that a chronic WET test failure (reduced weight) has a real chance of being the result of natural variability and not actual toxicity (i.e. another false positive). Other studies have also pointed out problems with variability and lack of measurable responses when using fathead minnow growth as the endpoint for toxicity tests (Mayer *et al.* 1986, Suter *et al.* 1987).

In summary, for the EPA criteria for chronic WET test acceptance/rejection based on the sub-lethal endpoints of reproduction or growth, this review indicates that up to 22% of chronic WET tests would be rejected for inadequate performance of controls with respect to *Ceriodaphnia* reproduction and up to 22% for fathead minnow growth. Since acceptance of chronic WET test results for any particular effluent relies on passing each of these criteria, this rejection rate would be additive, resulting in the potential for the rejection of up to 44% of all chronic WET tests due to inadequate control performance for sub-lethal endpoints. In addition, this analysis strongly suggests that some effluent test failures, when measured as reduced reproduction or reduced growth, could be the result of natural variability among the test organisms and not effluent toxicity.

2.2.2 Mortality Endpoint - Survival

Unlike the sub-lethal parameters of reproduction or growth, survival would be expected to exhibit a truncated, one-tailed distribution resembling half of a bell-curve (Wilson and Bossert 1971, McNaughton and Wolf 1973). Since survival cannot be greater than 100%, the majority of individuals would be expected to exhibit high survival with fewer numbers at the low end of the range of survival.

Survival data for the 217 chronic WET test controls were plotted for *Ceriodaphnia* and exhibited a range of 0 to 100 % survival (Fig. 3). A relationship resembling a truncated one-tailed distribution can be observed. Given these data, the mean survival for *Ceriodaphnia* for a 7-day test in toxin-free reconstituted laboratory water would be 88%, with a standard deviation of 22 (Fig. 3). For a one-tailed distribution, the t-value for the 95% confidence level becomes 1.64 instead of the 1.96 used in 2-tailed distributions (Snedecor and Cochran 1967). Using the mean $\pm$ 1.64 SD, the normal range of control survival for *Ceriodaphnia* is from 52% to 100% survival.

Using standard statistical procedures, *Ceriodaphnia* controls that perform within this range of survival would be adequate for test acceptance. EPA's criteria of 80% survival for control organisms is above the lower bound of the range of normal performance, but would invalidate only 6% of the controls that still fall within the mean $\pm$ 1.64 SD (Fig. 3). The survival (mortality) endpoint appears to be considerably more robust than the sub-lethal endpoint of *Ceriodaphnia* reproduction.

When the 235 chronic WET test controls are plotted for fathead minnow survival, a truncated one-tailed curve is again observed (Fig. 3). The mean survival for fathead minnows using these data for 7-day chronic test controls would be 91% survival with a standard deviation

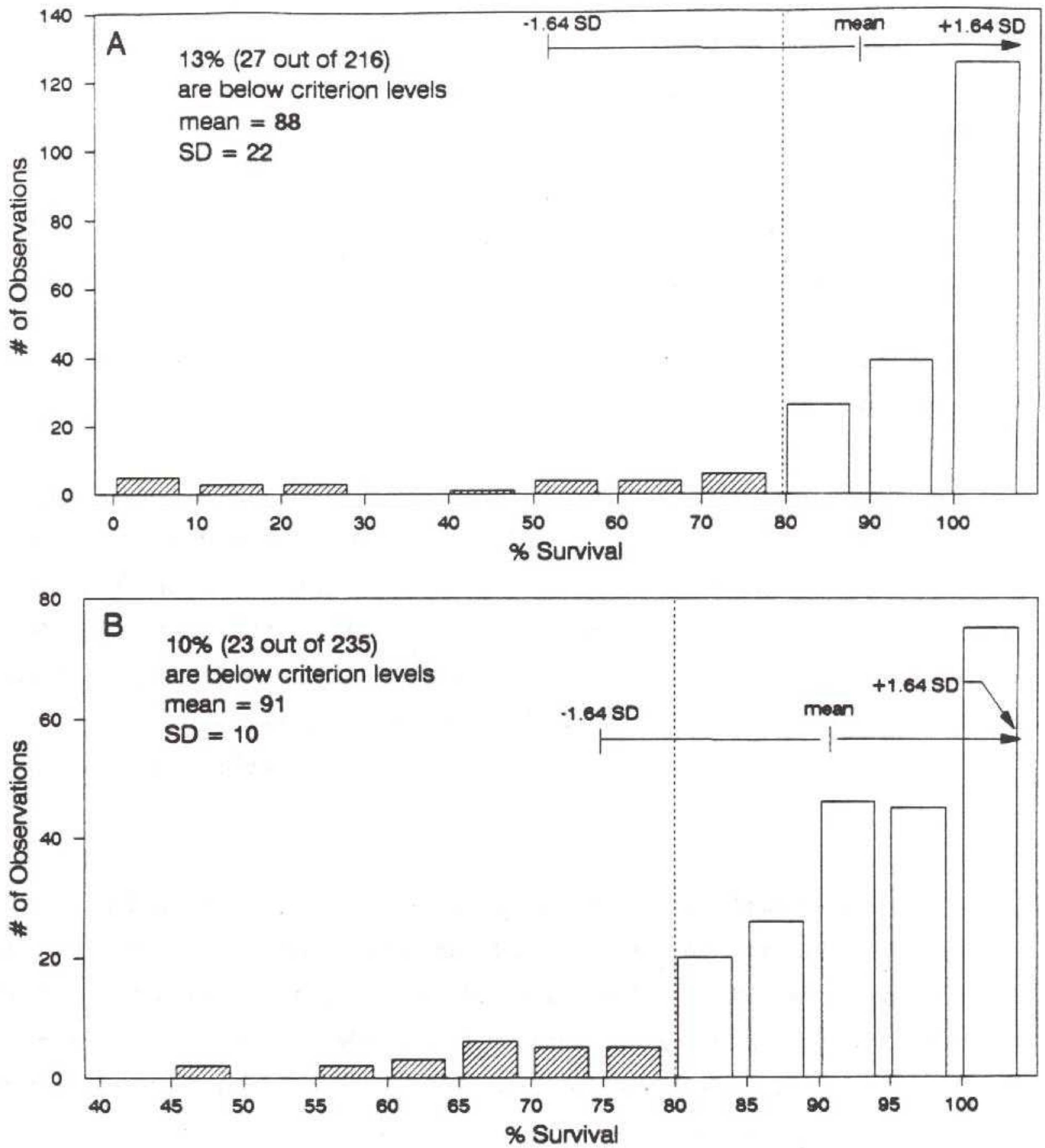


FIGURE 3: Mean survival (% alive) for *Ceriodaphnia* (A) and fathead minnows (B) from control organisms exposed only to effluent-free reconstituted laboratory water.

of 10 (Fig. 3). Using the mean $\pm$ 1.64 SD as described above, the normal range of control survival for fathead minnow larvae appears to be 78% to 100%.

Using standard statistical procedures, fathead minnow controls that perform within this range are adequate for test acceptance. EPA's criteria of 80% survival for fathead minnows falls just above the lower bound of the range of normal performance for fathead minnow survival. This analysis would indicate that with regards to fathead minnow survival, EPA control performance criteria of 80% would cause only 2% of the tests to be rejected that still fall within the natural range of variability inherent in fathead minnow survival. As with *Ceriodaphnia*, the survival (mortality) endpoint for fathead minnows is considerably more robust than the sub-lethal endpoint of growth.

In summary, for the EPA criterion of survival for chronic WET test acceptance/rejection, this analysis indicates that few (up to 8%) chronic WET tests would be falsely rejected for inadequate performance of controls with respect to *Ceriodaphnia* or fathead minnow survival.

Using survival as an endpoint in the chronic toxicity test appears to have a significant advantage over the use of sub-lethal endpoints: the survival parameter is considerably more robust than the sub-lethal endpoints of reproduction or growth. The expected rate of false rejection of either of the tests for inadequate control survival is 6% or less, compared to the rejection rates of 22% for *Ceriodaphnia dubia* reproduction and 22% for fathead minnow end-of-test weight.

2.2.3 Biological Variability - Implications to Test

The analysis of biological variability in control metrics for the chronic WET test shows that no two *Ceriodaphnia* or fathead minnows would necessarily be expected to exhibit equivalent responses, especially with sub-lethal endpoints. This is naturally expected when using living organisms. A number of studies have shown that these test organisms, especially the daphniids, will exhibit a considerable range in reproduction, growth and survival; even under controlled laboratory conditions (Anderson and Norberg-King 1991, Belanger *et al.* 1989, Cooney *et al.* 1992a,b, Cowgill 1987, DeGraeve and Cooney 1987, Frank *et al.* 1957, Frank 1957, 1960, Kraus and Kornder 1987, Murdoch and McCauley 1985, Patterson *et al.* 1992, Pickering 1988, Woltering 1984).

The implication of variability in the response of seemingly identical organisms is that the chronic WET tests cannot be adequately controlled in the classical sense of managing variables in experimental design. It is evident that, even with the strict test protocols presently required by EPA, natural variation in reproduction and growth preclude determination of toxicity with an acceptable degree of confidence. Average *Ceriodaphnia* reproduction under control conditions in reconstituted laboratory water can vary from 2 to 38 offspring (at the 95% confidence interval) and fathead end-of-test weight can vary from 0.02 to 0.84 mg/individual (see Table 1). The variability resulting from the wide range of observed responses has also been reported by a number of other researchers in the field of aquatic ecotoxicology (Anderson and Norberg-King 1991, Dhaliwal and Dolan 1991, Dorn and Rodgers 1989, Hall and Borton 1987, Kraus and Kornder 1987, Parkhurst *et al.* 1992, Parkhurst and Mount 1991, Warren-Hicks and Parkhurst 1992).

TABLE 1: Control performances for survival, reproduction (# of offspring) and/or growth for *Ceriodaphnia dubia* and fathead minnows from 7-day chronic toxicity tests (CV = Coefficient of variation). Data accumulated from 34 laboratories (see text for explanation).

Species	# of tests	mean % Surv.	S.D.	CV (%)	# of Tests	Offspring/ Weight(g)	S.D.	CV (%)
Ceriodaphnia dubia	217	88	22	25	214	20	9	46
Fathead Minnow	235	91	10	12	230	0.43	0.21	49

EPA protocols assure that only tests with controls above criteria will be used. Given the data presented above, it is apparent that controls will at times have organisms that perform below these criteria simply as a result of exhibiting the normal range of performance for that species. The potential rejection rates presented earlier should be added together to give an overall test rejection rate based on the survival and sub-lethal endpoints. The additive nature of the probabilities is the result of EPA's present protocols, which specify that an indication that any of the four metrics (*Ceriodaphnia dubia* survival, *Ceriodaphnia dubia* reproduction, fathead minnow survival, or fathead minnow end-of-test weight) is lower than control performance criteria, require rejection of the test. The sum of probabilities of a test being rejected due to inadequate control performance based on existing quality control protocols, is 52% (Table 2). In effect, there are four ways to reject a test, and each of the four ways to fail can represent the result of normal biological variability that is not adequately controlled under the existing protocol. The only way to pass the test is for all four metrics to match control performance closely, without regard to whether the result for each metric is within the normal range of variation for the species.

TABLE 2: Probability that an effluent test series will be rejected due to inadequate performance of control organisms.

Metric	Probability (%)
Ceriodaphnia dubia survival	6
Fathead minnow survival	2
Ceriodaphnia dubia reproduction	22
Fathead minnow end-of-test weight	22
Total probability of test rejection	52%

2.2.4 Coefficient of Variation for Control Organisms

The coefficient of variation (CV) for control performance by *Ceriodaphnia* and fathead minnows confirms that considerable variation is present, especially in the sub-lethal metrics of reproduction and end-of-test weight (Table 1). The CV for survival of organisms exposed only to reconstituted laboratory water is 12% for fathead minnows, but increases to 25% for *Ceriodaphnia*. For *Ceriodaphnia* reproduction, the CV increases further to 46% and for fathead minnow weight the CV is 49%. The coefficient of variation observed for the sub-lethal endpoints appears to be relatively high for a "controlled" laboratory test. In fact, these levels are more normally associated with the type of variability observed in natural populations, such as benthic invertebrates in streams (Canton and Chadwick 1988).

Similar high coefficients of variation have been reported for chronic WET test statistics, such as LC 50's and NOEC's (Anderson and Norberg-King 1991, DeGraeve *et al.* 199, 1992, Parkhurst 1992, Warren-Hicks and Parkhurst 1992), but this is the first time that such significant variability has been reported for control organisms that have only been exposed to effluent-free, reconstituted laboratory water. This analysis would indicate that much of the variation observed by other researchers in chronic WET tests is due to the inherent variability of the biological test organisms used rather than differences in test execution, sensitivity to toxicants, or other test-induced factors.

While the chronic WET test has apparent high variability, the EPA believes that this variability is no greater than that observed in "recognized" analytical chemical tests (EPA 1991a, DeGraeve *et al.*1992). Yet, a review of the EPA Discharge Monitoring Report Quality Assurance (DMR QA) Studies 11 and 12 shows that both acute and especially chronic WET tests have considerably higher CV's than the majority of standard water chemistry analyses (Table 3). This review used data from EPA's DMR QA program in which EPA sends samples of known quality to laboratories for analyses. In the case of WET tests, a highly toxic reference compound is used in the test dilution series. All the *Ceriodaphnia* WET test results had CV's of 30% or higher (Table 3). Fathead minnow tests also had CV's generally over 25% (Table 3). Of the other 30 analytes (including metals, nutrients and other parameters), only total

Table 3: Comparisons of analytical precision - Analysis of U.S. EPA's Discharge Monitoring Report quality assurance studies (DMR QA) Numbers 11 and 12.

	ANALYTE	COEFFICIENT of VARIATION DMR #11	DMR #12	SPIKE LEVEL DMR11	DMR12
METALS	Arsenic	7%	11%	200	100 ug/L
	Beryllium	8%	10%	130	100 ug/L
	Cadmium	6%	6%	190	250 ug/L
	Chromium	7%	7%	410	830 ug/L
	Cobalt	5%	5%	180	110 ug/L
	Copper	4%	4%	730	790 ug/L
	Iron	5%	5%	1000	1500 ug/L
	Lead	8%	5%	48	550 ug/L
	Manganese	3%	4%	920	809 ug/L
	Mercury	10%	10%	3.5	5.3 ug/L
	Nickel	5%	4%	430	740 ug/L
	Selenium	10%	19%	86	14 ug/L
	Zinc	8%	5%	110	360 ug/L
NUTRIENTS	Ammonia	8%	10%	4	2.5 mg/L
	Nitrate	7%	7%	10	15 mg/L
	Kjeldahl Nitrogen	9%	9%	25	34 mg/L
OTHER PARAMETER	Chem. Oxy. Demand	8%	11%	65	11 mg/L
	5-day BOD	17%	16%	41	36 mg/L
	Cyanide	11%	11%	.53	.61 mg/L
	Chlorine (TRC)	46%	17%	.11	.44 mg/L
	Phenolics	32%	19%	.015	.207 mg/L
	pH	1%	1%	5.6	9.4 s.u.
	Suspended Solids	11%	11%	24	25 mg/L
CHRONIC TOXICITY	Minnow-Survival (NOEC)	34%	39%	3.1	32.5 %conc
	Minnow-Growth (NOEC)	50%	50%	3.2	25.0 %conc
	Minnow-Growth (IC25)	41%	41%	4.7	46.4 %conc
	C.daphnia-Survival (NOEC)	50%	31%	3.5	15.6 %conc
	C.daphnia-Reprod. (NOEC)	50%	34%	2.2	15.0 %conc
	C.daphnia-Reprod. (IC25)	50%	45%	2.7	12.8 %conc
ACUTE TOXICITY	Minnow-Mortality (LC50)	29%	9%	5.6	69.2 %conc
	C.daphnia-Mortal (LC50)	40%	42%	5.1	19.9 %conc

residual chlorine and phenolics had CV's over 20%. The majority of parameters had CV's 10 times less than those for WET tests. Review of EPA DMR QA data shows that WET test are considerably more variable than standard water chemistry analyses.

This should not be surprising when the differences between chemical analyses and toxicity tests are considered. An analogy that compares toxicity testing using biological organisms to chemical testing using electronic instruments may be useful.

Most modern chemical analyses are performed using electronic instruments. Chemical analysis is preceded by intricate and detailed preparations designed to assure that the analytical instrument is working properly. The working parts of the instrument, particularly detector elements and hoses that carry the sample, are cleaned and flushed with sample. Operating temperatures, pressures, and flow rates are carefully adjusted. Electronic calibration is performed to assure that the circuits are operating within specified parameters, and the instrument is tuned to minimize signal noise. Next the instrument is sample calibrated. A series of solutions containing known concentrations of the analyte are analyzed. The chemist looks both for the expected results, and for a linear response that shows that when the concentration of analyte doubles, the signal from the instrument doubles. If this linear "dose response" is not obtained, the instrument is not operating properly and is adjusted accordingly. Finally, known solutions from two separate sources are tested to make sure that no systematic error has biased the system. When all these tests have been completed successfully, the chemist begins to analyze unknown samples, while repeating the analysis of known solutions every fifth sample or so to assure that the instrument is still tuned and calibrated and giving a proper response.

By contrast, when biological organisms are used for toxicity testing, each biological organism is an analytical instrument. They are instruments that cannot be tested, tuned or calibrated. Each instrument is disposable; it is used once and discarded. Instrument calibration in toxicity tests consists of the unproven assumption that all fathead minnows and all water fleas are essentially identical clones of one another, so that when individuals are exposed to similar conditions, they will react in a similar manner.

Unfortunately, as was observed in the wide range of reproduction and end-of-test weight, the assumption of genetic sameness clearly is not true. This fact was demonstrated by observing the performance of test species with respect to the sub-lethal test endpoints (weight gain for fathead minnows, reproduction for *Ceriodaphnia dubia*) in reconstituted laboratory water that is used for conducting control tests. Reconstituted laboratory water is a chemically defined medium. Test organisms are selected from common stock, representing a single genetic line. Temperature, the daily light/dark cycle, food, and the frequency and volume of feeding are the same for all organisms. Even with these precautions, a *Ceriodaphnia dubia* may produce anywhere from only 2 young to over 38 young during a defined time period; while one fathead minnow may weigh anywhere from 0.02 grams at the end of the test to over 0.80 grams. Genetic variability among biological organisms is inescapable.

To put it in terms of analytical chemistry, the genetic variability inherent in biological organisms constitutes a significant "background signal noise" that confounds the sub-lethal endpoints of the chronic WET test. The statistical power of the chronic WET test must be raised so that it is able to distinguish relatively small toxic effects from background signal noise, i.e. reduce the "signal-to-noise" ratio.

It is commonly observed that the dose response requirement is violated in evaluating the sub-lethal endpoints of the chronic toxicity test at low concentrations of effluent (DeGraeve *et al.* 199, 1992, Warren-Hicks and Parkhurst 1992). This is indicative of the inability of the test to distinguish small effects from background noise. It is comparable to the difficulty instrumentation has when an analyte is near the detection limit of the instrument. The result can again be false positive indications of toxicity where no toxicity really exists.

2.3 IMPACT OF VARIABILITY ON TEST PRECISION

Test protocols for chronic WET tests depend greatly on randomization to reduce the influence of biological variability (Weber *et al.* 1989). This includes choosing organisms at random for placement in test containers, placement of test containers at random in the environmental chambers and replacing containers in the chambers at random following the daily renewals/counts. However, randomization efforts can generally only be effective if a large number of replicates are utilized (Elliott 1977). The general recommendation for statistical design is a minimum of 30-50 replicates to minimize variability (Snedecor and Cochran 1967, Elliott 1977). The *Ceriodaphnia* tests use 10 replicates per treatment (60 organisms total) and the fathead minnow tests only use 3-4 replicates per treatment (180-240 organisms total); far short of that which may be needed to avoid recording false positives resulting from background "noise."

2.3.1 Number of Tests for Desired Precision

The standard approach used to assess the affect of the variability observed in studies such as the chronic WET test is to determine the number of repeated tests that are needed in order to achieve an acceptable level of confidence in the results (Elliott 1977). Biological organisms are variable and this must be accounted for in the determination of significance of results. This is the premise behind EPA's use of an Impairment Concentration percentage set at 25% (IC25) (Norberg-King 1988). In other words, an impairment would have to reach 25% of the sample population before it would be a measurable impact. Less than that would only likely be measuring biological "noise".

Using the 25% figure as a starting point, it was determined how many tests would need to be run to achieve a level of confidence of 25%, i.e., to ensure that the 95% confidence level is within 25% of the mean. This number was calculated based on the summary statistics in Table 1 and the equations in Elliott (1977). From these calculations it was determined that one test would be needed to achieve a 95% confidence interval that is within 25% of the mean for fathead minnow survival and three tests would be needed for *Ceriodaphnia* survival (Table 4). The increased variability observed in the sublethal parameters is reflected in the substantially greater number of tests needed to achieve the same level of confidence (Table 4). If the desired precision is increased to allow a 95% confidence interval that is only within 10% of the mean, the number of tests needed increases dramatically (Table 4), for both the survival and sub-lethal endpoint.

TABLE 4: Number of tests required to achieve a desired level of precision (as represented by sampling errror being within a certain % of the population mean) at the 95% and 75% confidence intervals for *Ceriodaphnia* and fathead minnow chronic WET test controls.
(Revised October 6, 1992)

Precision	*Ceriodaphnia* Survival	*Ceriodaphnia* Reproduction	Fathead minnow Survival	Fathead minnow Reproduction
95% C.I. within 25% of mean	3	13	<1	15
95% C.I. within 10% of mean	17	80	4	92
75% C.I. within 20% of mean	<1	7	<1	8

This desired level of precision can also be caculated to reflect a Practical Quantitation Level or PQL-type approach used in chemical laboratories, where three-quarters of the labortories should report values within 20% of the mean. This case would translate to a 75% confidence interval that is within 20% of the mean. At this level of precision, one test would be adequate for the survival endpoints for either *Ceriodaphnia* or fathead minnows. However, multiple tests would still have to be run to meet this level of precision for the sub-lethal endpoints (Table 4).

Conversely, the level of precision (i.e. what percentage of the mean is represented by the 95% confidence interval) that is achieved when conducting a certain number of tests can also be calculated from the data in Table 1 and the equations in Elliott (1977). The relationship between precision (as defined above) versus number of tests conducted is a hyperbolic-type curve (Figure 4). Note that even though permit compliance relies on the results of one test, the precision achieved with just one test cannot be calculated. The equation requires use of a t-value for n-1 degrees of freedom and there is no t-value for 0 degrees of freedom. Considering that failure of one test constitutes a permit violation, the unknown level of precision (with the sub-lethal parameters) achieved when only one test is conducted lends substantial doubt to the reality of that presumed "violation."

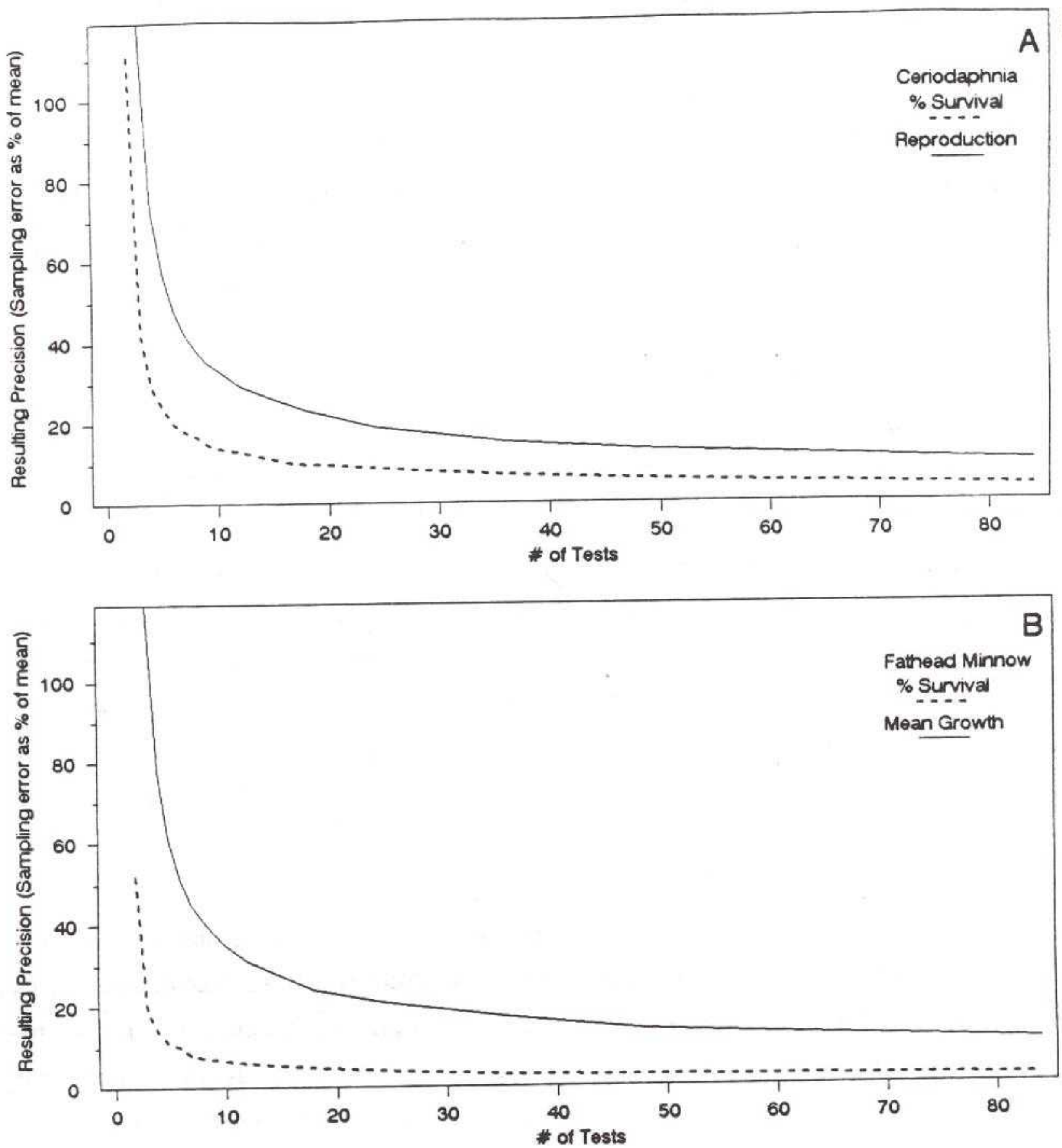


FIGURE 4: Calculated precision, defined as sampling error measured as the percent of the population mean represented by the 95% confidence interval, for survival and sub-lethal endpoints for *Ceriodaphnia* and fathead minnows.

Even when three tests are run, the precision for *Ceriodaphnia* reproduction is 114%. In other words, three *Ceriodaphnia* tests provide a 95% confidence interval for number of offspring/female that is within 114% of the mean. Precision for fathead minnow growth with three tests is similar at 121%. The observed relationships again point to the greater reliability of the survival endpoints. Precision for three *Ceriodaphnia* survival tests is a 95% confidence interval that is within 42% of the mean and three fathead minnow tests provide a precision of 19%. To achieve the level of precision discussed earlier of a 95% confidence interval that is within 25% of the mean, over 12 *Ceriodaphnia* tests and over 24 fathead minnow chronic tests would need to be run, if the sub-lethal endpoints are considered (Figure 4).

2.3.2 Compensation for Test Variability

The primary technique that could compensate for genetic variability is the use of a larger number of test organisms and replicates within each test. Within the full range of a biological species, it is possible to characterize "average" response for the species by measuring the response of a large number of randomly selected individuals. Unfortunately, the chronic WET testing protocols presently recommended by EPA use a relatively low number of test organisms. The present test uses only 180-240 fathead minnow larvae, representing 10 fish in each of only 3-4 replicates at 6 effluent dilution concentrations. For *Ceriodaphnia dubia*, only 60 organisms are used representing one neonate in each of 10 replicates at 6 effluent dilution concentrations.

A stark contrast to the chronic WET test protocol is provided by the pharmaceutical industry's bioassay tests used to determine the safety and efficiently of new drugs. Pharmaceutical bioassays are performed using plates of pure culture ("genetically identical") bacteria with hundreds of millions of bacteria per plate, and 50 or more plates (replicates) at each concentration. While bacteria display genetic variability in a manner similar to fathead minnow larvae and *Ceriodaphnia dubia*, the pharmaceutical tests provide considerably larger numbers of test organisms to make statistical tests powerful enough to quantify the smallest effects of the active substance being tested. The relatively small number of test organisms used in the WET testing protocols, particularly with *Ceriodaphnia dubia* at only 10 organisms per effluent concentration, cannot provide the statistical power required to reliably detect the presence or absence of toxicity, as measured by the sub-lethal endpoints of the chronic whole effluent toxicity test.

One means of compensating for biological variability would be to recognize the two-tailed distribution in the sub-lethal parameters of reproduction or growth. A two-tailed distribution should have control performance criteria that reject not only the poor performers at the low end, but also reject the super performers on the high end. In addition, two-tailed statistical tests should be used to determine whether sub-lethal endpoints in various effluent concentrations differ significantly from controls. Use of a two-tailed statistical test would reflect the fact that the test is equally susceptible to both false positive and false negative results and more realistically reflect the bell-curve distribution of reproduction and growth.

A second means of compensating for biological variability would be to apply the Practical Quantitation Limit (PQL) concept of compliance assessment for chronic WET test results, in the same manner that it is applied to compliance assessment for chemical testing. This PQL test calibration approach for chronic WET testing should be acceptable, just as it is for water chemistry, because EPA has stated repeatedly that it considers the sub-lethal endpoints of the chronic toxicity test to be equally reliable as chemical testing (EPA 1991a, DeGraeve *et al.* 1992).

A third modification would relate to controlling false positives (finding toxicity where none exists) by applying a rigorous dose-response requirement in interpreting the test. In other words, measured toxic effects must increase as effluent concentration increases. Thus, if a sub-lethal endpoint differs significantly from the control at an effluent concentration of 25 percent, but the sub-lethal endpoint does not differ significantly from the control at effluent concentrations of 50 or 100 percent, then the test should be judged invalid and discarded.

As noted earlier, EPA officials have stated unequivocally that a dose response is an essential characteristic of a valid chronic toxicity test:

"The dose response curve is the basis for the validity of a toxicity test. The control serves as the starting point from which the dose response is evaluated. If a dose response is not obtained, the toxicity cannot be inferred." (Norberg-King 1989)

"Finally, one of the mandatory requirements for a valid toxicity test, a predictable dose-response curve, cannot be observed. We would never accept analytical results from an instrument producing an abnormal standard curve. The predictable dose-response curve, that is increasing toxicity with increasing concentration, is the analogue of the analytical standard curve and is of equal importance to toxicity testing." (Mount 1990).

While the survival and sub-lethal endpoint cutoff criteria have been formalized in EPA chronic WET test guidance documents (Weber, *et al.* 1989), the dose-response requirement was not included in EPA's quality assurance protocols for the chronic WET test (although it is mentioned in the TSD, EPA 1991a). The omission of this critical control against false positive tests is a flaw in terms of the ability to use sub-lethal endpoints of the chronic toxicity test as an enforceable measure of permit compliance.

A fourth method of compensating for biological variability is to increase the statistical power of the test protocol by increasing the number of test organisms or replicates used in each test. For example, the number of *Ceriodaphnia dubia* organisms used at each effluent concentration should be at least equal to the number of fathead minnows. In addition, ideally, 30 replicates are required before the experimental design is said to based on a relatively "large" sample size (Snedecor and Cochran 1967).

2.4 VARIABILITY IS INHERENT TO TEST

Variability in the chronic test can be mistaken for sensitivity or questionable laboratory competence. *Ceriodaphnia* is usually considered the most sensitive species, when in fact the previous discussion shows that it may only be the most biologically variable species. Test variability is often attributed to laboratory competence (DeGraeve *et al.* 1992a,b). However, the results of over 200 controls from 34 different laboratories show that variability is not the result of any one laboratory's performance, but rather natural biological variation in test organisms. In fact, EPA recognizes the variable nature for test involving biological organisms; otherwise they would not have required replicates in order to account for background noise.

2.4.1 Variability not Induced by Laboratory Performance

Some have argued that biological variability is irrelevant because the test procedures identify it, eliminate it or control for it. The key assumption in this position is that any residual observed variability is being "induced" by something the laboratory did (or didn't do) to the samples. Since all samples are handled identically, it is further assumed that when control organisms fail to perform adequately (presumably due to poor test execution), and the test is aborted, any risk of biased results among effluent-exposed organisms was also neutralized when the test was abandoned and re-initiated.

All of this would be true <u>if</u> variability was induced by the laboratory and occurred uniformly among treatment groups. However, there is no evidence to support that assumption. Consider the fact that if this assumption were true, then a number of results would have to follow:

1) If a laboratory "mishandled" a control, thereby causing inadequate survival or reproduction/growth, then all containers/beakers in the entire chronic WET test should exhibit the reduced survival or reproduction (growth) noted in the control at the end of the test. But, this is not the case. A beaker by beaker analysis shows that one beaker with below criteria performance is generally the exception, with the other beakers performing at or above performance criteria (DeGraeve *et al.* 1989a,b, Chadwick & Associates, Inc. unpubl. data, ERC Environmental, Aqua Terra Technologies, and Riverside WWTF unpubl. data).

2) If the failure of the control to meet EPA performance criteria is a result of laboratory competence, then the accompanying effluent dilution series should also uniformly fail EPA performance criteria since they would be subjected to the same laboratory competence as the controls. This does not happen. Review of the over 50 chronic WET tests run by Chadwick & Associates, Inc. during the UAA (see Vol. 2, Appendix F) show that every time some aspect of control performance did not pass EPA performance criteria, one or more of the dilution series had metrics that did pass criteria.

3) If failure of controls is a result of laboratory competence, then simultaneously run controls should also fail since they are subjected to essentially similar practices and procedures. Again, this is not the case, as the review of the over 50 chronic WET tests run during the UAA shows (see Vol. 2, Appendix F).

4) If control performance is a result of laboratory competence, then those labs with better performance (i.e. higher mean reproduction) should exhibit less variability. This is not the case, as can be seen when reviewing the data from the different labs summarized in Table 1 (unpubl. data). While the statistics of reproduction or growth vary between labs, the standard deviation (i.e. variability) remains relatively consistent.

2.4.2 EPA Recognizes the Presence of Variability

This review is not the first to recognize that variability exists in the chronic WET test. EPA's *Technical Support Document for Water Quality-based Toxics Control* (TSD) discusses analytical variability in the chronic WET test at Section 5.7.6. The discussion points out that all environmental monitoring includes analytical variability, and there is no way to determine whether a reported value is higher, lower, or the same as the "true" value (EPA 1991a).

EPA states that one way to account for analytical variability would be to adjust the permit limit by adding or subtracting the analytical variability, but concludes that this approach is inappropriate. Adding analytical variability might have the effect of making the limit "less protective" of the water quality objective. The reverse of this argument, that subtracting analytical variability from the calculated limit could be overly conservative, especially in light of the other conservative factors used to develop permit limits, can not be made because of the way EPA structured the chronic WET testing protocols. The existing protocols are structured in a manner that protects against false negatives (finding no toxicity where toxicity exists), but does not protect against false positives (finding toxicity where toxicity does not exist). EPA's criteria make the probability of a false negative result near zero. Yet, as was concluded earlier, the probability of a false positive result may be high.

EPA has concluded that it is not possible to correct a permit limit to account for analytical variability and that the competing tendencies for overprotection and underprotection would balance out in the long run (EPA 1991a).

In response to public comments on the revised TSD, EPA provided further clarification of its position on enforcement of chronic WET permit limits. In the Responsiveness Summary (EPA 1991c), which addressed comments on the TSD, EPA provided the following response to Comment 12 on Chapter 2 (page 11 of the Responsiveness Summary):

> "The allowable frequency for criteria excursions should refer to true excursions of the criteria, not to spurious excursions caused by analytical variability or error. In evaluating data on chemical concentrations ***or toxicity units***, it is desirable to subtract the analytical error log variance from the observed log variance in order to arrive at the true log variance contributing to criteria excursions." [Emphasis added]

This amplification of EPA's position makes it clear that EPA does not intend that permit limits should result in enforcement actions as a result of apparent violations caused by analytical variability, either for chemical-specific numeric limits or for chronic WET test limits.

However, the response provided by EPA in the Responsiveness Summary contradicts the position expressed in the TSD (EPA 1991a,c). The TSD states that it is not possible to correct test results for variability; yet the Responsiveness Summary specifies a method for correcting test results for variability. In fact, the ability to correct test results to account for variability is in itself a function of the ability to quantify variability. The method is to subtract the variability inherent in the analytical method from the variability of the effluent, in order to partition natural biological variability from effluent variability. The ability to quantify analytical variability is determined by the statistical power of a test, which is primarily related to the number of test cases. In chronic WET testing, the number of test cases is equivalent to the number of test organisms exposed to each effluent concentration.

Because EPA considers it "inappropriate" to adjust permit limits to account for variability, but at the same time also considers it "inappropriate" for enforcement actions to be based on apparent excursions that may be caused by analytical variability or error, it is essential that analytical protocols incorporate the best science available to minimize variability and assure that test results reflect real conditions as accurately as possible. The chronic WET test must have high statistical power in order to discern small toxic effects from the background of natural genetic variability.

SECTION 3: CHRONIC TEST'S ABILITY TO RELIABLY PREDICT IMPAIRMENT TO AQUATIC ECOSYSTEMS

3.1 ABILITY OF CHRONIC WET TESTS TO PREDICT TOXICITY

The primary factor limiting the ability of chronic WET tests to predict impacts to the aquatic biota of receiving waters is the fact that the background variability observed in the sub-lethal endpoints of reproduction and growth is too large to adequately discern true toxicity. The analysis of control performance presented earlier shows that as currently structured, the chronic WET test cannot reliably predict the presence, absence or level of toxicity present in effluents with an acceptable degree of certainty. If no reliable measure of effluent toxicity is possible, no reliable prediction of impacts on receiving streams is possible.

A number of other constraints limit the ability of laboratory toxicity tests to predict effects on natural ecosystems, including:

1) Effluent samples may not represent in-stream water chemistry conditions.
2) Single-species tests may not represent in-stream aquatic biology conditions.
3) Test species may not be representative of native fauna.
4) The frequency and duration of exposure in the laboratory can not represent in-stream conditions.
5) Behavior activity that test organisms would exhibit in-stream is inhibited under laboratory conditions.

3.2 EFFLUENT SAMPLES CANNOT REPRESENT IN-STREAM CONDITIONS

Samples of effluent from "end of pipe" cannot represent water quality characteristics in the receiving stream, and yet it is this in-stream condition that the resident species inhabit. It is well known that toxicity of chemicals to aquatic life is dependant on a large number of factors, such as available dilution, completing by organic compounds and suspended material, antagonistic/synergistic effects, amelioration by hardness, different activities of chemicals at different temperatures and pH's, chemical and biological transformation of toxics, duration and frequency of toxics in effluent, and variable sensitivity of different organisms (Arthur *et al.* 1987, Buhler and Williams 1988, Cairns *et al.* 1975, Ingersoll and Winner 1982, Judy and Davies 1979, Lewis and Horning 1991, Moore and Winner 1989, Munkittrick *et al.* 1991, Pascoe *et al.* 1986, Paulauskis and Winner 1988, Thurston *et al.* 1981a,b, Winner 1986).

EPA documented the phenomenon that certain common water quality conditions could ameliorate toxic effects of many chemical constituents through a variety of mechanisms which may not be specifically definable. In recognition of this possibility, the EPA included the "indicator species" procedure in its 1983 Water Quality Standards Handbook (EPA 1983). The indicator species procedure relies on water effects ratios (WER), which are calculated by comparing the toxicity of a chemical in receiving water to the toxicity of that chemical in reconstituted laboratory water. When the chemical is demonstrably less toxic in receiving water than in reconstituted laboratory water, the ratio of toxicity results can be used to scale up water quality objectives in receiving water. This method relies on the assumption that some factor or combination of factors present in the receiving water lessens the toxic impact of the chemical. EPA recently reaffirmed its commitment to the use of WERs by developing more comprehensive guidance on the use of the technique for establishing heavy metal water quality criteria (EPA 1992).

Laboratory test conditions cannot represent in-stream conditions for aquatic life. In natural systems, organisms use a number of strategies when confronted with adverse water quality. This includes avoidance of toxic plumes, assimilation and biotransformation of toxics, acclimation to toxics, and natural selection for populations resistant to the toxics (Black and Birge 1980, Buhler and Williams 1988, Cairns and Niederlehner 1987, Holcombe *et al.* 1976, 1979, Little *et al.* 1990, Nakamura 1986).

The implication of these facts, together with the previous discussion, is that chronic WET tests conducted on effluents before they mix with receiving waters, even assuming that the chronic WET tests are completely reliable, cannot predict the environmental effects of these effluents once they mix with the receiving waters. The high variability associated with the sub-lethal endpoints of the chronic WET test compounds this problem. As a practical matter, the results of chronic WET testing, as measured by the sub-lethal endpoints and carried out as specified in NPDES discharge permits, cannot be considered to have any demonstrated ability to predict the occurrence of impairment to natural aquatic communities.

3.3 TEST CONDITIONS CANNOT REPRESENT IN-STREAM CONDITIONS

An important factor often neglected in the discussion of chronic WET tests and the variability in the range of response observed in the sub-lethal endpoints is the imposition of "test stress" on the organisms. This is a common factor considered in other types of testing. Researchers often assume that test stress is accounted for by running a control series along with the effluent dilution series. However, as noted earlier, these controls still exhibit considerable variation that may, in part, be a result of test stress, induced through handling activities. These could include daily renewals of water in the test chambers, intermittent feeding, counting and removing neonates, etc.

Within the established methodology of the 7-day chronic WET tests, additional variation is induced by the specific culturing techniques and slight variations of test conditions that can occur even when fully complying with EPA test protocols. The presence and amount of variance produced by each individual test, culture, or analytical parameter is unknown, but may not affect controls and effluent dilutions equally and thus may skew the measure of toxicity.

3.3.1 Culturing procedures

The guidelines within the EPA manual for culturing and rearing of test organisms are fairly specific (Weber *et al.* 1989). However, a range of conditions (i.e. lighting, temperature, food quality) are allowed, recognizing time constraints, equipment availability, and other laboratory specifics. Minor changes in these parameters can result in additional test variability.

One source of variance is the amount and quality of food fed to the "stock" organisms. For example, the type, size, and nutritional quality of *Artemia* cysts, the food for fathead minnow larvae, may vary between individual vendors as well as within lots from a single supplier (Weber 1991). Even if the quality of the cysts appears sufficient, it is possible for a batch of *Artemia* to hatch out at a size too large to be available as food for the first few days of the fathead minnow early life stage test. Variances due to food quality/source are also prevalent in invertebrate culture procedures. Belanger *et al.* (1989) and Cowgill *et al.* (1985a,b) found that the EPA recommended diet of 1:1 *Selenastrum*:YCT (yeast, Cerophyll®, trout chow) can produce inaccuracies when testing substances high in heavy metals because of the potentially significant amounts of metals in the YCT mixture. In addition the quality of the algae can vary according to undetected bacterial contamination and differences within the allowed range of cell concentration (3.0 to 3.5 x 10^7 cells/mL of algal concentrate). These factors would affect both the controls and effluent dilutions causing additional variability, adversely affecting the ability to statistically measure toxic effects.

3.3.2 Dissolved Oxygen Compensation

Dissolved oxygen (DO) conditions in the fathead minnow test include a compensatory measure that is put in place when the DO level falls (Weber *et al.* 1989). If during an effluent test, DO falls below 40% of saturation, calculated according to air pressure and elevation, artificial aeration must be added. This aeration is to account for effluents which are highly oxyphilic, (i.e., high biological oxygen demand) such as those with the presence of large amounts of organic materials (e.g. POTW discharges). However, this protocol could result in correcting for oxygen depletion "after the fact" (i.e. after the less than 40% saturation is measured). Thus, the protocol includes the potential for mortality due to oxygen depletion rather than chemical toxicity before the problem is fixed. In addition, if the aeration is not carefully applied, too much turbulence can cause added stress to the organisms (Weber 1991). The additional energy expenditure spent on swimming may not allow the controls to achieve the 0.25mg mean 7-day end-of-test weight required for test acceptance. Weber (1991) also notes that aeration can also increase potential ammonia toxicity by altering pH levels. Conversely, aeration has been shown to reduce toxicity in other parameters, especially volatile organics (Belanger *et al.* 1988, Weber *et al.* 1989, Weber 1991).

3.3.3 Physical conditions

The physical parameters specified in the testing methodology (Weber *et al.* 1989) include such factors as temperature, amount of light, duration of light:dark cycle, exposure chamber size, and test solution volume. In comparison to the biological and chemical conditions, the physical conditions are generally more controllable, and thus variability should not be as severe. Yet, recently Chadwick & Associates noted a slight, but significant, increase in mean reproduction by *Ceriodaphnia* simply by shifting the light:dark cycle to start earlier in the day (5:00 am instead of 8:00 am). Apparently, this early time gave the organisms a head start in reproduction prior to handling at the start of the work day and may relate to the handling stress noted earlier.

Another physical condition is the daily renewal of test water and the receipt of "fresh" test water every other day during the chronic test (Weber *et al.* 1989). Physically removing the "old" water and replacing it with "fresh" water adds a daily test stress to the organisms. This removal can also induce additional variability through the addition of water during the test that may not be of the same quality, due to some variation in effluent quality, storm events, and other episodes of short-term duration (Lewis *et al.* 1989, Stewart *et al.* 1990).

As noted before, while these factors will affect both control and effluent exposed organisms, each additional source of variability makes the test less able to reliably distinguish the presence or absence of toxicity.

3.3.4 Analytical procedures

EPA provides procedures for the analysis of toxicological results involving the statistical application of hypothesis testing and point estimation techniques (Weber *et al.* 1989). No-observed effect concentration (NOEC) and lowest observed effect concentration (LOEC) are calculated by hypothesis testing and analysis of variance with such procedures as the Dunnett's procedure or the Bonferroni t-test (used when a data set doesn't contain equal replicates). Other hypothesis testing techniques, Steel's Many One Rank and the Wilcoxin Rank Sum, are employed if the data set is not normally distributed or if there is heterogeneous variance.

Hypothesis testing procedures, such as Dunnett's, have often been criticized because the response levels are constrained to arbitrarily selected exposure levels in the test (e.g. 6.25%, 12.5%, 25%, 50% 75%, 100% effluent) and the test has difficulty handling test variability within each dilution (Berger and Ellgas 1991, Dantin 1991, DeGraeve *et al.* 1989a, Dhaliwal and Dolan 1991, Masters *et al.* 1991, Norberg-King 1988, Suter *et al.* 1987). Notably, Dunnett argued against the use of his statistical method for toxicity tests, emphasizing that rejecting controls for inadequate performance reduces the ability to detect false positives (Dunnett, C.W. 1973. Multiple Comparisons. pp 10. In: *Proceedings, 29th Annual Princeton Conference on Applied Statistics* - as cited in Masters *et al.* 1991)

Using point estimation techniques of statistical analysis, such as calculation of Impairment Concentration (ICp) has recently been promoted by EPA as an alternative to NOEC. These approaches are intended to allow a more reliable determination of the true effect level (Norberg-King 1988, DeGraeve *et al.* 1989a). However, this method is also biased in that it has apriori assumptions of 1) the presence of toxicity and 2) a defined dose - response relationship. This method can not recognize when biostimulation confounds test results (DeGraeve *et al.* 1989a).

Determination of toxicity should be independent of the statistical test used; yet the tests run for the UAA show that determination of toxicity is not independent of the test used (see Vol. 2, Appendix F). The impairment concentration of 25% (IC25) was originally chosen because it is supposed to give the same result as the NOEC calculated using Dunnett's procedure (Norberg-King 1988). However, during the UAA, there were numerous instances where the NOEC and IC25 statistical tests gave significantly different answers regarding the toxicity of the effluent. For example, the fathead minnow chronic WET test for Chino RPII effluent in August 1991 had an NOEC of 100% effluent, while the IC 25 was 17% effluent (UAA Vol 2, p. F-47).

The implication of inconsistencies in results from statistical tests run on the same data combined with the inability of toxicity test conditions to represent instream conditions accurately is that statistical significance in a toxicity test may have no relevance to biological significance instream.

3.4 TEST ORGANISMS ARE NOT REPRESENTATIVE OF SANTA ANA RIVER FAUNA

One of the common problems with single-species toxicity tests is that the test organisms are rarely representative of the native fauna of the receiving water and thus may not truly measure toxic effects on resident biota (Cairns 1983, 1984, 1986, 1992, Cairns and Pratt 1989, Cairns and Niederlehner 1987, Gray 1989, Maciorowski and Clarke 1980, Maltby and Calow 1989, Moore and Winner 1989). This can be especially true in a system like the Santa Ana River, with harsh physical habitat conditions and a naturally depauperate fauna (see Vol 2).

3.4.1 Fathead Minnows and *Ceriodaphnia* are not native

As was discussed in greater detail in UAA Volume 2 (pp 39-41), the only fish species native to the middle Santa Ana River are the Santa Ana sucker and the arroyo chub. Fathead minnows are not native and probably entered the basin in a "bait bucket" transfer. While fathead minnows are now present in the Santa Ana River, the similarities or differences from the native species in terms of response to effluents are not known (except to note that both fathead minnows and the two native species are thriving in an effluent dominated stream).

Ceriodaphnia dubia is not a resident species of the Santa Ana River, nor could it ever be. *Ceriodaphnia* is a planktonic, lake dwelling organism and would not be expected to be a stream resident since it cannot withstand current (Balcer *et al.* 1984, Brandlova *et al.* 1972, Dodson and Frey 1991, Pennak 1989). Of additional concern is the lack of basic ecological knowledge of *Ceriodaphnia dubia*. *Ceriodaphnia dubia* is a European species that may have entered the United States in *Daphnia magna* cultures, which is another European species (S.

Dodson, University of Wisconsin, personal communication). There remains considerable controversy regarding the taxonomy of this exotic species (Berner 1987, Weber 1991), which makes its use by EPA as the "standard" test organism surprising. EPA apparently conducted no formal research effort directed at scanning a large number of invertebrate species in order to identify the most appropriate test organism.

In any event, it is no more appropriate to use *Ceriodaphnia dubia* for toxicity testing in the Santa Ana River than it is to assign water quality objectives that are driven by cold-water species like trout that cannot exist in the Santa Ana River. *Ceriodaphnia dubia* would never be a significant component of the aquatic community of any flowing water, and it is not native to North America, much less the Santa Ana River basin. If it is necessary to assure that Santa Ana River water quality is protective of lake species, then it would be more appropriate to perform toxicity testing with a species like *Daphnia pulex*, which at least is known to exist within the watershed.

3.4.2 Species Give Conflicting Results

Ceriodaphnia provides conflicting results when chronic WET tests are conducted on sample splits simultaneously at different laboratories. A recent group of split sample tests was conducted by four different laboratories for the Chino Basin Municipal Water District (Chino Basin MWD unpubl. data). A total of five split tests were conducted over 9 months (Fig. 5). All the labs used EPA chronic WET test protocols for *Ceriodaphnia*. In only one case did the results of the *Ceriodaphnia* chronic WET test match between two laboratories. In the rest of the tests (4 out of 5 tests), the two labs gave substantially conflicting results of toxicity (or the lack thereof) and no one lab gave more consistent toxic responses (Fig. 5). This lack of repeatability has also been found in a recent study by the City of Riverside (unpub. data). In this case, effluent samples were sent simultaneously to three laboratories, which conducted 7-day chronic WET tests on *Ceriodaphnia dubia* using standard EPA protocols. The NOEC's based on the sub-lethal endpoint of *Ceriodaphnia* reproduction were 100% effluent for one laboratory, 50% effluent for the second laboratory, and 12.5% effluent for the third laboratory.

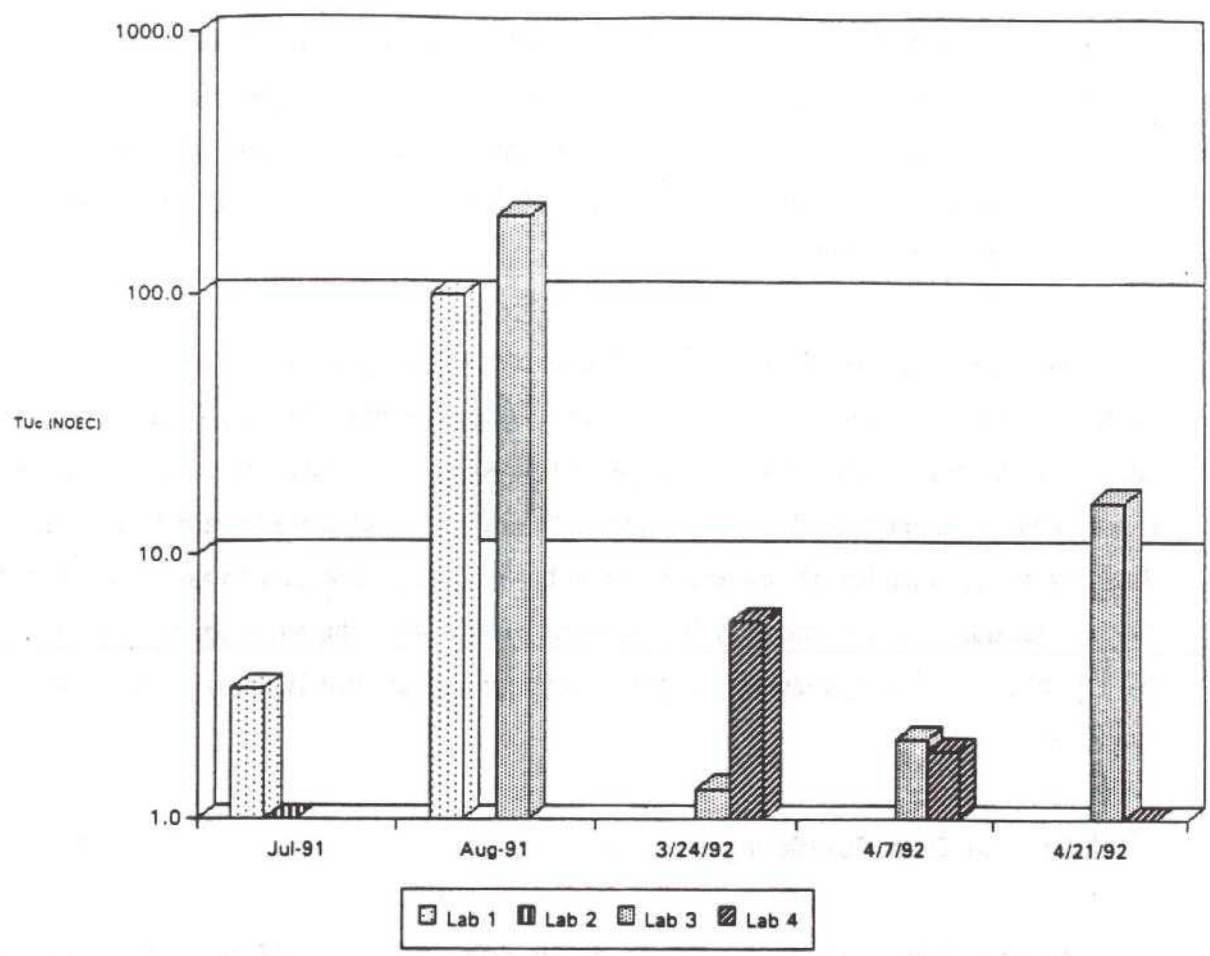


FIGURE 5: Split Chino Basin MWD RP-1 final effluent sample chronic WET test results, expressed as TUc, for *Ceriodaphnia* from four different laboratories (Chino Basin MWD unpubl. data).

Ceriodaphnia also gives conflicting results when compared to other cladoceran species. The literature would tend to indicate that *Ceriodaphnia dubia* is generally more variable and sensitive to culturing and test conditions than other cladocerans (Cowgill *et al.* 1985a, Kszos *et al.* 1992, Winner 1988). Thus, *Ceriodaphnia* results may not be comparable to other species with a longer history of use in toxicity tests or those species more likely to be representative of resident species in the Santa Ana River.

This concept was tested by running chronic WET tests simultaneously with three different cladocerans: *Ceriodaphnia dubia, Daphnia magna* and *Moina affinis* to evaluate the toxicity of effluent from a very highly treated, advanced waste treatment plant (Chino Basin MWD unpubl.

data). *Ceriodaphnia dubia* was used since it is the chronic WET test organism recommended by EPA (Weber *et al.* 1989). *Daphnia magna* is a common chronic WET test organism that is sometimes used in acute toxicity tests as well (Weber 1991). A related species, *Daphnia pulex*, has been found in nearby Southern California lakes (Chino Basin MWD unpubl. data). *Moina affinis* is a cladoceran found in the southwestern United States (Pennak 1989) and frequently exhibits populations blooms in Chino Basin MWD clarifiers.

When results of chronic WET tests are compared between species, it is evident that *Ceriodaphnia dubia* gives results that are considerably different from the other two cladoceran species (Fig. 6). In no case did the other species measure toxicity, whereas there was always some indication of toxicity (as reported in Toxic Units) by *Ceriodaphnia dubia*. In addition, *Ceriodaphnia* itself exhibited wildly different toxic responses between tests (Fig. 6), with no observed difference in effluent quality (Chino Basin MWD pers. commun.). These analyses indicate that *Ceriodaphnia* may be an inappropriate test species, especially for the Santa Ana River basin, because it does not show consistent responses to test conditions and would not be expected to be found in the Santa Ana River basin. Continued reliance on test results using *Ceriodaphnia dubia* may result in permit conditions, which are unnecessarily over-protective of the Santa Ana River system.

Another problem with the use of *Ceriodaphnia* and fathead minnows in the Santa Ana River is the conflicting indications of the presence or absence of toxicity when both species are used. Frequently through the UAA, significantly different chronic WET test results were found when comparing the sub-lethal endpoints for *Ceriodaphnia* and fathead minnows (see Vol. 2, Appendix F). This was also found in a recent study of the RIIX effluent conducted by the City of San Bernardino (unpubl. data). In this case, one site had an NOEC for the sub-lethal endpoint of end-of-test dry weight for fathead minnows of 100% effluent, while for *Ceriodaphnia* reproduction the NOEC was 18% effluent. Notably, for the survival endpoint the NOEC was 100% effluent for both species. At another site, this pattern reversed with an NOEC for fathead growth of 32% effluent, but an NOEC of 100% effluent for *Ceriodaphnia* reproduction (San Bernardino, unpubl. data).

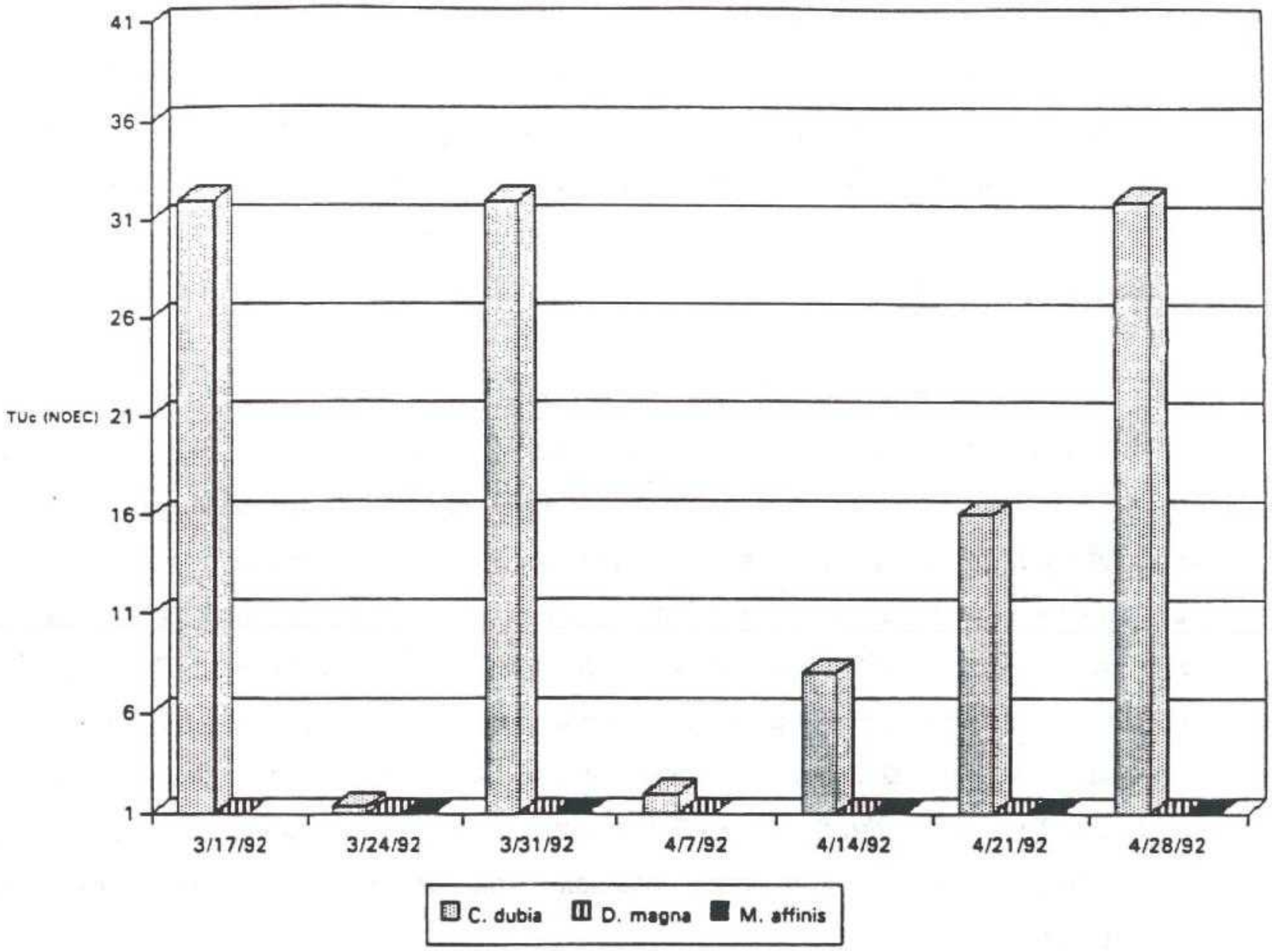


FIGURE 6: Split Chino Basin MWD RP-1 final effluent sample chronic WET test results, expressed as TUc, for *Ceriodaphnia dubia, Daphnia magna* and *Moina affinis* (Chino Basin MWD unpubl. data).

3.5 UAA PROVIDES A CASE STUDY

A primary conclusion of the UAA was that there was no evidence for impairment to aquatic life in Reach 3 of the Santa Ana River or in Chino Basin, when compared to the Santa Margarita River (UAA, Vol. 2, pp. 108-110). This conclusion was reached after collection and review of over 20,000 data points collected seasonally on the fish, invertebrate and algal populations as well as in-stream habitat and water quality, (see Vol 2 and 3). Despite evidence for non-impairment, chronic WET testing of effluents (Riverside and Chino RPI & RPII) and sample sites downstream (SAR 8 & 9, CC2) did show intermittent but inconsistent toxicity for either *Ceriodaphnia* or fathead minnows (Vol. 2, pp. 97 - 99). However, there was no correlation between the toxicity test results in these stream reaches and the corresponding indicators of biological integrity, especially with regards to species abundance and diversity. These conflicting results call into question the applicability of chronic WET tests using the sub-lethal endpoints in the prediction of impairment to the Santa Ana River ecosystem. Thus, a closer review of the basis for EPA's assertion that chronic WET test results reflect actual in-stream impairment is needed.

3.6 EPA FIELD VALIDATION STUDIES

Field validation studies conducted by EPA were designed to confirm whether chronic WET tests accurately predict impairment to receiving waters. These include the eight EPA validations studies (Complex Effluent Toxicity Testing Program - CETTP) conducted for EPA by Mount and Norberg-King (1985, 1986), Mount *et al.* (1984, 1985, 1986a,b,c) and Norberg-King and Mount (1986). In addition, in a recent review of this issue (EPA 1991), similar studies conducted by Birge *et al.* (1989), Eagleson *et al.* (1989) and Dickson, *et al.* (1992) were also included. Unfortunately, these studies are fundamentally flawed (Parkhurst *et al.* 1990). The primary deficiency is the lack of adequate statistical design and analysis to provide definitive answers to the research question (Parkhurst *et al.* 1990). Most of the results use only qualitative correlation to "prove" a relationship between measured chronic toxicity and instream impairment. Parkhurst *et al.* (1990) lists a number of critical factors not accounted for in these studies including habitat variability, flow events, sediment input, non-point sources, salinity, low temperature effects, un-monitored point sources, and dissolved oxygen deficits.

Our review of these studies has uncovered other serious flaws in the study designs. One major problem in the Skeleton Creek, Oklahoma, study (Norberg-King and Mount 1986) was the presence of an un-monitored wastewater treatment plant in the study reach. Another major flaw was the lack of consistent, predefined control sites for the various parameters. For example, again in the Skeleton Creek study, reference (control) stations were chosen to be the site which exhibited the least toxicity or highest number of taxa, regardless of whether that site was above or below effluent discharges (see Norberg-King and Mount 1986, p. 8-3). In addition, each parameter could have a different reference site, whether it be *Ceriodaphnia* toxicity, fathead minnow toxicity, benthic invertebrate species richness or fish species richness. Using this approach, control sites often occurred at sites downstream of the very effluents being tested. This design violates several basic tenets of ecological study design and selection of reference sites, including those in EPA's own Rapid Bioassessment Protocols (Plafkin *et al.* 1989). Using this design, stream sites upstream of all effluents could be measured as impaired, if the effluent had an enrichment effect.

In the other studies, sampling and analysis methods varied between years, reference sites changed between years, and sampling of effluents and stream sites were sometimes conducted over a month apart.

A consistent flaw that appears in all the CETTP studies is the failure to partition acute and chronic toxicity. The EPA's acute toxicity test has been shown to be a reasonably stable test, leading to an acceptable level of correlation between predicted and observed in-stream effects. However, the CETTP studies did not distinguish clearly between acute toxicity effects and chronic toxicity effects. As a result, EPA's claim that good correlation was observed between instream effects and effects predicted using the sub-lethal endpoints of the chronic WET test is invalid, since the CETTP study sites were specifically placed where acutely toxic effluents were known to be present. Acute and chronic toxicity were not partitioned in EPA's studies, resulting in multicolinearity between the two independent variables. Multicolinearity constitutes a major problem in EPA's analyses, because it is impossible to attribute effects observed instream to either acute or chronic toxicity with any degree of accuracy.

Another flaw that occurs throughout the CETTP studies is the failure to select sites randomly. All the sites tested during the CETTP studies were chosen on the basis of expected impairment (EPA 1991a). This mode of site selection constitutes inadequate experimental design. If expected impairment is a valid criterion for site selection, then steps should have been taken to assure that study sites encompassed a variety of sites where impairment was expected, where impairment was not expected, and where the level of impairment was unknown prior to the study. Random selection of sites ***over the entire range of expected results*** is essential for to valid study design. EPA's selection of study sites only where impairment is expected and not defining predetermined control sites introduces systematic error into the study, and biases the study toward a higher probability that chronic WET tests showing toxicity will be positively correlated to instream impairment. A proper study design would have included follow-up validation studies on rivers where no impairment was expected, where only chronic toxicity was expected, or where the level of impairment was unknown.

On the basis of these studies, EPA states that the probability of false positives is in the range of 12% (Brandis 1989, EPA 1991a). Because EPA's study design allowed acute and chronic toxicity to become confounded, a more appropriate way to state this observation is that even under optimum test conditions, a false positive rate of at least 12% is expected.

At this stage of review, there are enough study design deficiencies in the CETTP studies to lend serious doubt to the conclusions that chronic WET tests can reliably predict instream conditions. EPA relied heavily on a reanalysis of the CETTP and associated data conducted by Dickson *et al.* (1992), to disagree with the conclusions contained in the critical review prepared by Parkhurst *et al.* (1990). EPA continues to believe that chronic WET tests accurately predict impairment of receiving water (EPA 1991a). However, in their review of these CETTP studies, Dickson *et al.* (1992) acknowledge that the significant correlations found between toxicity and in-stream impairment were aided greatly by the presence of high ambient toxicity. They simply state that their analysis showed a strong relationship existed "in the data sets examined." With regards to other cases, they state:

> "This high level of ambient toxicity [in the data sets] assisted in making it possible to observe a relationship between ambient toxicity and community response. Where ambient toxicity is low or marginal...it will be more difficult to elucidate a relationship. Noise in the instream biological response variables and the confounding factors such as habitat and residual sediment toxicity may make observing a relationship difficult or impossible. For these reasons, it appears that the examined data sets, in spite of their inherent limitations, were appropriate for

the objective of evaluating the relationships between ambient toxicity and instream biological impact. If data sets were examined where ambient toxicity was low or marginal and no relationship was observed, the analysis would have resulted in an ambiguous interpretation. As in classical toxicology, dose and exposure are everything. The methods reported in this paper require that a toxicity gradient exists in the data sets, as is true for any method that examines the relationships between two sets of variables."

The chronic WET test procedure has not gone under any significant, critical per review. Nor has it been formally certified in standard methods (40 CFR 136) for use as a pass-fail test for NPDES permit compliance. Until the chronic toxicity testing protocols have been more thoroughly evaluated and validated, they cannot *reliably* predict the presence or absence of toxicity or reliably predict impairment to aquatic ecosystems.

SECTION 4: IMPLICATIONS FOR NPDES COMPLIANCE

From the results of the analyses described in this volume, it appears that considerable variability is inherent in evaluating the sub-lethal endpoints of the chronic WET tests due to biological factors. The variability observed in chronic WET test results has a number of practical implications for water quality program management. Variability affects the cost of monitoring, the reliability of results, the usability of results, the ability to respond to indications of toxicity, and the relationship among the three major elements of water quality program management (numeric criteria, toxicity testing, and biological criteria). Perhaps the most immediate implication is the manner in which variability in the sub-lethal endpoints of the chronic WET test affects the use of these tests in NPDES permits, where they are sometimes used as single-test "pass/fail" limits.

4.1 FALSE RESULTS REDUCE TEST UTILITY

4.1.1 Implications of Inadequate Control Performance

Ceriodaphnia dubia controls must produce an average of 15 young per female, fathead minnow controls must exhibit an average end-of-test weight of 0.25 grams, and controls for both species must exhibit at least 80% survival for a chronic WET test to be deemed acceptable. The combined expected rate of test rejection due to control failure in chronic WET tests may be as high as 52% percent. Using a conservative assumption that the actual rate of control failures may only be half this high, almost one out of four tests would still need to be re-run at the expense of the laboratory or the discharger. Based on an average cost of $2,000 for a two-species chronic WET test, dischargers required to perform monthly or biweekly testing could expect to spend and extra $12,000 to $24,000 per year for repeat testing, assuming that no positive test results are obtained, raising the annual cost for chronic WET monitoring from $24,000 to as high as $48,000.

In addition to the monetary costs of retesting, the value of lost data must be considered. Chronic WET tests require significantly longer completion times than chemical tests. Even in-house laboratories would have difficulty returning fully analyzed results in less than 15 days from the start of the test. Commercial laboratories can be expected to take up to 30 days to return official results that have been fully reviewed for quality control purposes. This means that, by the time the test results are obtained from the testing laboratory, the opportunity to re-sample in order to obtain toxicity information on an effluent during a specific time period has been lost. It is not possible to preserve additional volumes of sample for re-testing, because samples for toxicity testing cannot be preserved like samples for chemical testing (i.e. with acid), and the time required to obtain an initial test result would violate holding time requirements for substances like volatile organics, ammonia, nitrite and total residual chlorine. The variability of chronic WET testing greatly decreases the utility of this test and increases the costs of using chronic WET testing to confirm that specific treatment processes, pretreatment programs, or source control programs can be effective at removing chronic toxicity effects.

4.1.2 Implications of False Positive Results

The analysis presented earlier shows that the rate of chronic WET tests indicating toxicity where no toxicity exists (false positive results) may be high, causing chronic WET tests to indicate NPDES permit violations even though toxicity may not be present. This has serious implications to the discharger, who could be subject to fines or imprisonment associated with permit violations when a sub-par group of *Ceriodaphnia dubia* accidentally winds up in the wrong beaker. In such a case, low reproduction by *Ceriodaphnia dubia* would be evaluated as an indication of toxicity, when in fact it was simply a result of natural biological variability.

The implications of false positive results are magnified by the fact that EPA has rejected the approach developed by California and a number of other states that would reduce the number of permit limit violations generated by false positive results. California's approach required that more than one chronic WET test failure must occur before a permit limit is violated. However, EPA insists that each individual test result must be considered "independently enforceable." In fact, EPA has gone so far as to suggest that the Clean Water Act may require that the result of each effluent fractionation procedure performed during a TIE or TRE be considered independently enforceable (King 1992). Thus a discharger, particularly a municipal discharger, might have a record of apparent intermittent toxicity which can never be defined as to cause or source. The municipality is liable for penalties associated with permit violations which may be simply false positives. Moreover, case law has established the principle that when a permit limit based on a 30-day average is violated, the discharger is liable for 30 individual daily fines (Gwaltney 1987). Because chronic WET testing is seldom required more frequently than once per month, and daily civil penalties of up to $10,000 are authorized under the Clean Water Act, each chronic WET test failure carries a potential liability of up to $300,000 in fines.

It is not necessary that fines be assessed in order for chronic WET test failures to have adverse impacts on municipal dischargers. The mere existence of a record of permit violations increases insurance costs and degrades municipal bond ratings. In addition, EPA's guidelines for assessing penalties for permit limit violations include an examination of compliance history. If a discharger experiences a series of false positive results, this pattern could be construed to represent a history of noncompliance, leading EPA or state agencies to seek far higher penalties in enforcement actions than they would otherwise consider justified. The existence of a record of apparent toxicity test failures created by false positive chronic WET results also invites third-party citizen suits which the 1987 Clean Water Act amendments authorized.

In addition, dischargers in California are required to take "all reasonable steps" to eliminate toxicity. If the cause and source of toxicity cannot be identified, which is the case when apparent toxicity is really intermittent false positive results, then advanced waste treatment may be deemed the only reasonable option. Advanced waste treatment could mean extremely expensive processes like reverse osmosis or activated carbon filtration.

Finally, false positive chronic WET tests can result in the construction of unneeded, expensive advanced waste treatment processes like reverse osmosis or activated carbon filtration. Industrial pretreatment programs can only deal with industrial effluents, while significant levels of over-the-counter pesticides, automotive products (antifreeze, gasoline, motor oil), and household disinfectants in municipal sewage come from domestic sources. Source control programs rely on public education and voluntary cooperation. Pretreatment and source control cannot be depended upon to produce timely results for a municipality that has experienced a series of false positive tests. A city may feel that there is no choice other than the extreme option of constructing these costly advanced waste treatment systems, or moving to a zero-discharge option by diverting effluent discharges out of surface waters entirely. And after the advanced waste treatment processes or complete reuse systems have been constructed and made operational, chronic WET testing may continue to show the same rate of failure.

4.3 INABILITY TO INTERPRET OR COMPLETE TIE/TREs

The California Inland Surface Waters Plan (ISWP) establishes procedural requirements that dischargers must follow when chronic WET testing indicates the presence of effluent toxicity. Dischargers are immediately required to initiate an accelerated schedule of chronic WET testing until three consecutive tests indicate the presence of toxicity, or until three consecutive tests indicate the absence of toxicity. If three consecutive chronic WET results indicate the presence of toxicity, the discharger is required to undertake a Toxicity Identification Evaluation (TIE), using the procedures described in EPA guidance (EPA 1991b), aimed at identifying the chemical classes of the responsible toxicants. The TIE should lead to a Toxicity Reduction Evaluation (TRE) in which sources of the responsible toxicants are determined and toxicity is eliminated by appropriate measures that may include, but are not limited to, advanced waste treatment or source control.

Both the TIE and TRE procedures are based on the results of further chronic WET testing on effluent samples that are "known" to be toxic. The TIE methodology requires that a number of manipulations be performed on an effluent sample to eliminate the presence of specific classes of chemicals, then running chronic WET tests on the series of effluent fractions created by these manipulations, plus the unmodified sample, in order to determine by elimination what classes of chemicals contribute to effluent toxicity.

Chronic WET test variability can confuse the interpretation of the TIE results, and may make it impossible to proceed to the TRE. The theory on which TIE operates is that only one or a few of the effluent fractions created through chemical manipulation will show the absence of toxicity. These non-toxic fractions represent the class of chemicals responsible for toxicity. False positive results could obscure these critical observations, while false negatives could produce a false indication. Only a single false positive result would be required to send the investigation down a false trail. Only a single failed control would invalidate a series of up to 20 chronic tests running concurrently. A TRE based on the result of a TIE that was influenced by variable toxicity tests would likely prove fruitless.

4.4 CONFLICT WITH OTHER INDICATORS OF WATER QUALITY

EPA's water quality management program is based on a three-part requirement of proof for environmental integrity. The three parts are chemical-specific numeric criteria, toxicity testing, and biological criteria.

The three-part proof of environmental integrity has evolved since 1987. Prior to the adoption of the 1972 Clean Water Act, water quality assessments were based on biological surveys. At this time, it was often felt that biological surveys were problematic, because they were time consuming, labor intensive, and could often not provide definitive conclusions on water quality. In order to eliminate these deficiencies, the water quality management program implemented by EPA in 1976 relied on chemical-specific numeric criteria. By 1987 it was apparent that EPA could not reasonably generate criteria for the thousands of chemicals that may occur in surface waters. In addition, chemical-specific numeric criteria are limited in that they cannot provide protection against the synergistic effects of several chemicals which alone might not be toxic, but which could act in concert to produce toxicity. For this reason, toxicity testing procedures were developed to supplement chemical-specific criteria and strengthen the water quality management program.

By 1989, EPA became concerned that even the dual tests of chemical-specific numeric criteria and toxicity testing might not be adequate to determine environmental integrity. This concern was raised by observations that some surface waters did not appear to support the level of diversity or abundance of aquatic organism that they should, based on comparisons to similar streams. Thus, the EPA determined that a third element, biological criteria, was necessary as an evaluation of biological integrity. Not only should waters be free from specific chemicals above specified levels, and at the same time free from toxicity as measured by biological toxicity testing protocols, but waters should be equally productive as similar "reference" waters. Compliance with biological criteria is to be assessed using biological surveys.

A problem that has recently arisen is the issue of conflict among the three key indicators. The Santa Ana River provides an example where a surface water does not exhibit concentrations of chemicals that exceed chemical-specific numeric objectives and where the aquatic community is measurably more diverse and more abundant than an unpolluted reference river, yet there are sporadic chronic WET test results that indicate the presence of toxicity. Toxic effects cannot be corroborated either through chemical testing or though biological survey. Some tie-breaking mechanism has to be used to determine whether a toxicity problem genuinely exists in the Santa Ana River. Given the expected rate of false positive results, and considering the evidence provided by chemical testing and the biosurvey comparisons between the Santa Ana River and a reference stream, it seems reasonable to assume that at least some of the indications of toxicity in the Santa Ana River may be spurious, particularly in Reach 3 and Chino Creek.

The issue of conflict among indicators of environmental integrity raises serious questions for the process of setting public policy for water quality management. It is not reasonable to make public policy as if all three indicators were independently valid, when the evolution of the three-part system resulted from the need to correct the inadequacies of individual indicators. This issue must be resolved in order for rational water quality programs to be developed.

4.5 UNENFORCEABLE EFFLUENT LIMITATIONS

The ISWP requires chronic WET testing for all municipal wastewater treatment facilities. The ISWP states that dischargers are considered to be in violation of their discharge permits when three consecutive chronic WET tests indicated the presence of toxicity. However, EPA Region IX has objected to this portion of the ISWP.

Determining compliance with permit limits is a technical decision. Case law has established that technical decisions must meet certain minimum standards of rationality, that there must be adequate accounting for various factors including analytical variability, that there must exist notice of what action will result in a violation, and that consideration should be given to the question of whether a limit can reasonably be met given available technology (Pifher and Egan 1992). The genetics of both *Ceriodaphnia dubia* and the fathead minnow are poorly understood. Random mortality, *Ceriodaphnia dubia* reproductive success, and fathead minnow weight gain make the test difficult to control. *Ceriodaphnia dubia* nutrition is poorly understood. As one researcher stated: "There seems to be a large body of folklore, black magic, and experience in the biomonitoring community that may solve these nutritional problems, but the use of these techniques is inappropriate for a nationwide testing scheme" (Grimes 1987). This statement can be applied to several aspects of the chronic test, including the daily light regime, maintaining adequate rates of control survival and fathead minnow weight gain.

Given the wide range of response naturally present in *Ceriodaphnia* and fathead minnow populations, it would be difficult to determine the true presence or absence of toxicity using the results of chronic WET tests. Chronic WET test controls have an inherent variability of up to 49% as measured by the coefficient of variation (Table 1). For this reason, it might be impossible to use the sub-lethal endpoints of the chronic toxicity test to prove that violation of a permit limit has occurred, under the rules of evidence that must be satisfied in judicial proceedings. In addition to the legal question, Dr. Donald Mount, EPA's senior research scientist who headed the development of chronic WET testing protocols, has stated that he would not consider the result of a single toxicity test result to be conclusive evidence of effluent toxicity (Mount 1989). Strict application of the sub-lethal endpoints of the chronic toxicity test as "pass/fail" permit limits is inappropriate, both legally and technically.

4.6 CONTINUOUS LIABILITY

The variability inherent in the chronic WET test makes it impossible for any discharger to maintain consistent compliance with California's state-wide objective of 1 TUc, regardless of effluent quality. The objective of 1 TUc means that any indication of toxicity, no matter how slight, violates the objective. The variability of the chronic WET test virtually guarantees that false positive results will produce an intermittent record of toxicity. Under the provisions of the ISWP, three consecutive chronic WET results indicating no toxicity are required for the closure of a toxicity incident. Municipal dischargers in the Santa Ana basin are required to conduct monthly chronic WET testing, generating a minimum of 12 chronic WET results annually. Even under the conservative assumption that variability will result in only 15% false positives, dischargers are almost certain to have at least two positive chronic WET results per year, even if their effluents are actually free from toxicity. The result of two false positive tests per year is added testing, added monitoring expense, loss of public confidence in the quality of the effluent, and (potentially) the installation of unneeded advanced waste treatment processes or diversion of all effluent discharges from effluent-maintained streams. It is reasonable to assume that the variability of the chronic WET test will raise significant questions about regulation, enforcement, due process, and the role of scientific evidence in the Santa Ana River basin. The terms relating to WET testing and permit limits which EPA insisted be placed in the current permits issued to Santa Ana River dischargers remove much of the flexibility that the ISWP would have provided.

SECTION 5: CONCLUSIONS AND RECOMMENDATIONS

5.1 CONCLUSIONS

The chronic WET test, based on current guidance and procedures, should not be used as a pass-fail system to determine compliance with NPDES permits. The test cannot reliably distinguish the presence or absence of toxicity and does not reliably predict impairment to aquatic ecosystems.

Test organisms exhibit a high level of biological variability particularly in growth and reproduction. This background variation is not adequately controlled by normal randomization techniques and may cause deductive errors regarding the presence of toxicity.

High levels of variability may be mistaken for sensitivity in the test organism. The statistical methods specified by EPA are insufficiently robust to partition normal background variability from toxicity-induced impairments without the use of significantly larger sample sizes (i.e. more replicates).

Effluent characteristics are not representative of in-stream conditions. Test conditions are not representative of field conditions. And, test organisms are not representative of species found in the Santa Ana River. Chemical tests and biological surveys show that the sub-lethal endpoints of chronic WET tests frequently predicted toxicity where none was present. Therefore, chronic toxicity test results can provide only a superficial indicator of the health of the river system.

Continued reliance on current guidance and enforcement policies will be counterproductive to the cause of environmental protection. Unacceptable rates of false positive and false negative results will reduce test utility, discourage Toxicity Identification and Reduction Evaluations, create unreasonable liabilities, and divert scarce resources to unnecessary and ineffective treatment strategies.

Despite the EPA's commitment to "independent applicability," chronic WET test failures can not establish an enforceable permit violation until the test becomes more stable and reliable. It is likely that, until the test protocols are improved, all enforcement actions based on biomonitoring results will be the subject of litigation. The 1TUc water quality objective, as currently defined, established a standard of perfection for POTW performance which cannot be attained realistically because of the variability inherent in the test. Under the circumstances, it is reasonable to require that the test used to assess such performance be equally free of error.

5.2 RECOMMENDATIONS

In order to make the test more reliable and robust, it is necessary to establish a Method Detection Limit (MDL) and a Practical Quantitation Limit (PQL) for the chronic WET test. Just as with chemical testing, an MDL would indicate when the presence or absence of toxicity could be reliably demonstrated (99% confidence) based on results of a single test. A PQL would establish a level at which the degree of toxicity, measured as TUc, could be reliably quantified (99% confidence).

To establish an MDL & PQL, certain changes to the toxicity testing protocols would be appropriate. For example:

5.2.1 Use the chronic toxicity test to monitor plant performance trends and trigger Toxicity Identification Evaluations only. Discontinue using the test to assess permit compliance except where a clear and persistent pattern of toxicity is demonstrated with appropriately large samples.

5.2.2 Improve the statistical procedures by reducing the alpha-level from $p<0.05$ to $p<0.01$ for all hypothesis testing within the chronic toxicity protocols. And, revise the test statistics to use a two-tailed hypothesis test for sub-lethal effects in order to compensate for enrichment effects in diluted effluent. These changes will provide a higher level of statistical confidence (99%) in the test results.

5.2.3 Establish data validation criteria for systematically reviewing reported toxicity test results before DMR certification. Validation review should assure that lab procedures conformed to all required protocols, that control organisms performed to minimum levels recommended by EPA, that a dose-response was observed in any test that indicated toxicity, and that effluent exposed organisms performed significantly worse than "normal" for the species population before toxicity is inferred.

5.2.4 Modify control performance criteria for sub-lethal parameters to account for two-tailed distribution by rejecting both poor performers and super-performers. This will reduce the rate of false positives by excluding controls that are far above average performance.

5.2.5 Use only mortality data to make determinations of significant effects. Data from growth and reproduction measurements should only be used to corroborate conclusions drawn from the survival data.

5.2.6 Increase the number of replicates and test organisms in order to increase statistical power.

5.2.7 Substitute a resident invertebrate from the Santa Ana River basin for the *Ceriodaphnia dubia* in the toxicity testing protocols. This will provide a more reliable and consistent indicator of potential impairment to the river.

5.2.8 Move the point of compliance from end-of-pipe to in-stream. As with using resident species, this will provide a more accurate and reliable means of identifying actual threats to beneficial uses.

5.2.9 Compensate for inherent biological variability by adjusting the formula for calculating TUc. Use (100-Population Coefficient of Variation)/NOEC for each species. This would significantly reduce the probability that normal biological variability would be mistaken for the presence of toxicity.

5.2.10 Use an SSO translator mechanism to increase the TUc objective based on evidence of no downstream impairments and results from Toxicity Identification Evaluations. Average ambient TUc demonstrates what is required to protect existing uses when no impairment is indicated in-stream.

5.2.11 The Regional Board should establish certification requirements for all laboratories conducting NPDES permit-related biomonitoring. The Regional Board should compile a regional database of population performance characteristics for each species used in the chronic toxicity test protocols and conduct periodic "blind" studies using reference toxicants and reconstituted laboratory water to establish limits of test variability.

Some of these recommendations can and should be implemented immediately by the dischargers. Other suggestions will require formal authorization from the Regional Board. A few may even require State Board approval.

5.3 SUMMARY

The EPA would like to use the results from toxicity tests to determine compliance in much the same manner as chemical tests are presently used. There are, however, several key differences.

Chemical test results are easier to validate because the instrumentation can be calibrated. Method detection limits (MDL) and practical quantitation (PQL) limits can be established. Repeated measures can be used to determine confidence levels associated with concentration determinations.

Because the chronic toxicity test relies on living organisms as the "instrument" of detection, it is impossible to calibrate the test in the traditional sense. Given inherent biological variability, it is critical that any inference of toxicity be based on sufficiently large samples and robust statistical procedures.

However, precisely because the test procedure relies on samples and statistics, there will always be some risk of false positive and false negative indications. While normally such indications are inconsequential, such is not the case where the water quality objective requires absolute perfection.

The recommendations given above are intended to make the imperfections of the test less significant. But, it will remain impossible under any circumstances to reliably meet a 1TUc requirement when determinations of compliance are based on current protocols and practices.

LITERATURE CITED

Anderson, S.L. and T.J. Norberg-King. 1991. Precision of short-term chronic toxicity tests in the real world. *Environmental Toxicology and Chemistry* 10:143-145.

Arthur, J.W., C.W. West, K.N. Allen, and S.F. Hedtke. 1987. Seasonal toxicity of ammonia to five fish and nine invertebrate species. *Bulletin of Environmental Contamination and Toxicology* 38:324-331.

Balcer, M.D., N.L. Korda, S.I. Dodson. 1984. *Zooplankton of the Great Lakes: A Guide to the Identification and Ecology of Common Crustacean Species*. University of Wisconsin Press, Madison.

Belanger, S.E., J.L. Farris, and D.S. Cherry. 1988. Reduction in organic effluent static acute toxicity to fathead minnows by various aeration techniques. *Environmental Pollution*. 50:189-210.

Belanger, S.E., J.L. Farris, and D.S. Cherry. 1989. Effects of diet, water hardness, and population source on acute and chronic cooper toxicity to *Ceriodaphnia dubia*. *Archives of Environmental Contamination and Toxicology* 18:601-611.

Berger, R., and W.E. Ellgas. 1991. *Determining the Biological Significance of Toxicity Test Results*. East Bay Municipal Utility District, Oakland.

Berner, D.B. 1987. Taxonomy of *Ceriodaphnia (Crustacea: Cladocera)* in the U.S. Environmental Protection Agency Cultures. EPA/600/S4-86/032.

Black, J.A., and W.J. Birge. 1980. *An Avoidance Response Bioassay for Aquatic Pollutants*. Research Report No. 123. Univ. Kentucky Water Resources Research Institute, Lexington.

Birge, W.J., J.A. Black, T.M. Short, and A.G. Westerman. 1989. A comparative ecological and toxicological investigation of a secondary waste water treatment plant effluent and it's receiving stream. *Environmental Toxicology and Chemistry* 8:437-450.

Brandlova, J., Z. Brandl, and C.H. Fernando. 1972. The Cladocera of Ontario with remarks on some species and distribution. *Canadian Journal of Zoology* 50:1373-1403.

Brandis, R. 1989. Memorandum to Permit Branch Chiefs, EPA Regions I-X, dated May 10, 1989. Subject: Studies confirming the relationship between toxicity tests and receiving stream effects. Regulatory Development Branch, Washington DC.

Buhler, D.R., and D.E. Williams. 1988. The role of biotransformation in the toxicity of chemicals. *Aquatic Toxicology*, 11:19-28.

Cairns, J. Jr. 1983. The perspective of an aquatic ecologist. pp. 27-30. In: Graber, R.C. (Exec. Dir.). *Proceedings of the Seminar of Development and Assessment of Environmental Standards*. American Academy of Environmental Engineers, Annapolis.

Cairns, J. Jr. 1984. Are single species toxicity tests along adequate for estimating environmental hazard? *Environmental Monitoring and Assessment* 4:259-273.

Cairns, J. Jr. 1986. What is meant by validation of predictions based on laboratory toxicity tests? *Hydrobiologia* 137:271-278.

Cairns, J. Jr. 1992. Paradigms Flossed: The coming of age of environmental toxicology. *Environmental Toxicology and Chemistry* 11:285-287.

Cairns, J. Jr., A.G. Heath, and B.C. Parker. 1975. The effects of temperature upon the toxicity of chemicals to aquatic organisms. *Hydrobiolgia* 47:135-171.

Cairns, J. Jr., and B.R. Niederlehner. 1987. Problems associated with selecting the most sensitive species for toxicity testing. *Hydrobiologia* 153:87-94.

Cairns, J. Jr., and J.R. Pratt. 1989. The scientific basis of bioassays. *Hydrobiologia* 188/189: 5-20.

Canton, S.P. and J.W. Chadwick. 1988. Variability in benthic invertebrate density estimates from stream samples. *Journal of Freshwater Ecology* 4:291-297.

Cooney, J.D., G.M. DeGraeve, E.L. Moore, W.D. Palmer, and T.L. Pollock. 1988. Effects of food and water quality on culturing and toxicity testing of *Ceriodaphnia dubia*. Electric Power Research Institute. EPRI EA-5820.

Cooney, J.D., G.M. DeGraeve, E.L. Moore, B.J. Lenoble, T.L. Pollock, and G.J. Smith. 1989. Effects of environmental and experimental design factors on culturing and testing of *Ceriodaphnia dubia*. Electric Power Research Institute. EPRI EN-6468.

Cooney, J.D., G.M. DeGraeve, E.L. Moore, W.D. Palmer, and T.L. Pollock. 1992a. Effects of food and water quality on culturing of *Ceriodaphnia dubia*. *Environmental Toxicology and Chemistry* 11:823-837.

Cooney, J.D., G.M. DeGraeve, E.L. Moore, B.J. Lenoble, T.L. Pollock, and G.J. Smith. 1992b. Effects of environmental and experimental design factors on culturing and toxicity testing of *Ceriodaphnia dubia*. *Environmental Toxicology and Chemistry* 11:839-850.

Cowgill, U.M. 1987. Critical analysis of factors affecting the sensitivity of zooplankton and the reproducibility of toxicity test results. *Water Research* 21:1453-1462.

Cowgill, U.M., I.T. Takahashi, and S.L. Applegath. 1985a. A comparison of the effect of four benchmark chemicals on *Daphnia magna* and *Ceriodaphnia dubia-affinis* tested at two different temperatures. *Environmental Toxicology Chemistry* 4:415-422.

Cowgill, U.M., H.W. Emmel, and I.T Takahashi. 1985b. Inorganic chemical composition of trout food pellets and alfalfa used to sustain *Daphnia magna* Straus. *Bulletin of Environmental Contamination and Toxicology* 14:890-896.

Dantin, D.D. 1991. The reliability of the 7-day early life stage toxicity test for the prediction of chronic responses of two heavy metals, cadmium and copper, using the Japanese Medaka (*Oryzias latipes*). M.S. Thesis, University of Southwestern Louisiana, Lafayette.

DeGraeve, G.M., J.D. Cooney. 1987. *Ceriodaphnia:* An update on effluent toxicity testing and research needs. *Environmental Toxicology and Chemistry* 6:331-333.

DeGraeve, G.M., J.D. Cooney, T.L. Pollock, N.G. Reichenback, J.H. Dean, M.D. Marcus, and D.O. McIntyre. 1989a. Precision of the EPA seven-day fathead minnow larval survival and growth test: Intra- and interlaboratory study. Electric Power Research Institute. EPRI EA-6189.

DeGraeve, G.M., J.D. Cooney, B.H. Marsh, T.L. Pollock, and N.G. Reichenbach. 1989b. Precision of the EPA seven-day *Ceriodaphnia dubia* survival and reproduction test: Intra- and interlaboratory study. Electric Power Research Institute. EPRI EN-6469

DeGraeve, G.M., J.D. Cooney, D.O. McIntyre, T.L. Pollock, and N.G. Reichenbach, J.H Dean and M.D. Marcus. 1991. Variability in the performance of the seven-day fathead minnow (*Pimephales promelas*) larval survival and growth test: an intra- and interlaboratory study. *Environmental Toxicology and Chemistry* 10:1189-1203.

DeGraeve, G.M., J.D. Cooney, B.H. March, T.L. Pollock, and B.G. Reichenbach. 1992. Variability in the performance of the 7-d *Ceriodaphnia dubia* survival and reproduction test: An intra- and interlaboratory study. *Environmental Toxicology and Chemistry* 11:851-866.

Dhaliwal, B.S., and R.J. Dolan. 1991. *Aquatic Toxicity Data Interpretation and Application in Regulatory Compliance.* Central Contra Costa Sanitary District, California.

Dickson, K.L., W.T. Waller, J.H Kennedy and L.P. Ammann. 1992. Assessing the relationship between ambient toxicity and instream biological response. *Environmental Toxicology and Chemistry* 11:1307-1322.

Dodson, S.I., D.G. Frey. 1984. Cladocera and Other Branchiopoda. pp 723-786. In: Thorp, J.H. and A.P. Covich (eds.). *Ecology & Classification of North American Freshwater Invertebrates*. Academic Press, Inc., New York.

Dorn, P.B. and J.H. Rodgers, Jr. 1989. Variability associated with identification of toxics in national pollutant discharge elimination systems (NPDES) effluent toxicity tests. *Environmental Toxicology and Chemistry*. 8:893-902.

Eagleson, K.W., D.L. Lenat, L.W. Ausley and F.R. Winborne. 1989. Comparison of measured instream biological responses with responses predicted using the *Ceriodaphnia dubia* chronic toxicity test. *Environmental Toxicology and Chemistry* 9:1019-1028.

Elliott, J.M. 1977. *Statistical Analysis of Samples of Benthic Invertebrates*. Scientific Publ. No. 25, 2nd ed. Freshwater Biological Association, Ambleside, England.

Frank, P.W. 1957. Coactions in laboratory populations of two species of *Daphnia*. *Ecology* 38:510-519.

Frank, P.W. 1960. Prediction of population growth form in *Daphnia pulex* cultures. *The American Naturalist* 94:357-372.

Frank, P.W., C.D. Boll, and R.W. Kelly. 1957. Vital statistics of laboratory cultures of *Daphnia pulex* DeGeer as related to density. *Physiological Zoology* 30:287-305

Gray, J.S. 1989. Do bioassays adequately predict ecological effects of pollutants? *Hydrobiologia* 188/189:397-402.

Grimes, M.M. 1987. The impact of EPA's biomonitoring policy on POTWs. *Journal of the Water Pollution Control Federation* 59(8):755-760.

Gwaltney of Smithfield, Ltd. v. Chesapeake Bay Foundation, 791 F.2d 304 (4th Cir. 1986), vacated 108 S.Ct.376 (1987).

Hall, T.J., and D.L. Borton. 1987. Chronic effluent bioassays and their limitations in characterizing effluents. pp. 257-260. In: *TAPPI Proceedings 1987 Environmental Conference.*

Holcombe, G.W., D.A. Benoit and E.N. Leonard. 1976. Long term effects of lead exposure on three generations of brook trout (*Salvelinus fontinalis*). *Journal of the Fisheries Research Board of Canada* 33:1731-1741.

Holcombe, G.W., D.A. Benoit and E.N. Leonard. 1979. Long term effects of zinc exposure on brook trout (*Salvelinus fontinalis*). *Transactions of the American Fisheries Society* 108:76-87.

Ingersoll, C.G., and R.W. Winner. 1982. Effect on *Daphnia Pulex* (De Geer) of daily pulse exposures to copper or cadmium. *Environmental Toxicology and Chemistry* 1:321-327.

Judy, R.D. Jr., and P.H. Davies. 1979. Effects of calcium addition as $Ca(NO_3)_2$ on zinc toxicity to fathead minnows, *Pimephales promelas*, Rafinesque. *Bulletin of Environmental Contamination and Toxicology* 22:88-94.

King, E. 1992. Letter from Ephriam King, Chief of EPA's NPDES Program Branch, to Martha E. Rudolph, Assistant Attorney General, Natural Resources Section, State of Colorado, date June 25, 1992.

Kraus, G. and S. Kornder. 1987. Experience with EPA chronic bioassay methods. pp. 261-268. In: *TAPPI Proceedings 1987 Environmental Conference*.

Kszos, L.A., A.J. Stewart and P.A. Taylor. 1992. An evaluation of Nickel toxicity to *Ceriodaphnia dubia and Daphnia magna* in a contaminated stream and in laboratory tests. *Environmental Toxicology and Chemistry* 11:1001-1012.

Lewis, M.A., W.S. Eckhoff, and J.D. Cooney. 1989. Impact of an episodic event on the toxicity evaluation of a treated municipal effluent. *Environmental Toxicology and Chemistry* 8:825-830.

Lewis, P.A., W.B. Horning II. 1991. Differences in acute toxicity test results of three reference toxicants on *Daphnia* at two temperatures. *Environmental Toxicology and Chemistry* 10:1351-1357.

Little, E.E., R.D. Archeski, B.A. Flerov, and V.I. Kozlovskaya. 1990. Behavioral indicators of sublethal toxicity in rainbow trout. *Archives of Environmental Contamination and Toxicology* 19:380-385.

Liu, E.H. 1992 Santa Ana River use attainability analysis - Comments on effluent and stream toxicity. *USEPA* Region IX, San Francisco.

Maciorowski, H.D. and R. McV. Clarke. 1980. Advantages and disadvantages of using invertebrates in toxicity testing. pp. 36-37. In: Buikema, A.L. and J.Cairns, Jr., (eds.), *Aquatic Invertebrate Bioassays*. ASTM STP 715. American Society for Testing and Materials.

Maltby L., and P. Calow. 1989. The application of bioassays in the resolution of environmental problems; Past, present and future. *Hydrobiologia* 188/189:65-76.

Masters, J.A., M.A. Lewis and D.H. Davidson. 1991. Validation of a four-day *Ceriodaphnia* toxicity test and statistical considerations in data analysis. *Environmental Toxicology and Chemistry* 10:47-55.

Mayer, F.L. Jr., K.S. Mayer, and M.R. Ellersieck. 1986. Relation of survival to other endpoints in chronic toxicity tests with fish. *Environmental Toxicology and Chemistry* 5:737-748.

McNaughton, S.J., and L.L. Wolf. 1973. *General Ecology*. Holt, Reinhart and Winston Inc., New York.

Moore, M.V., and R.W. Winner. 1989. Relative sensitivity of *Ceriodaphnia dubia* laboratory tests and pond communities of zooplankton and benthos to chronic copper stress. *Aquatic Toxicology* 15:311-330.

Mount, D.I. 1989. Comments made during a videotaped public workshop on whole effluent toxicity testing presented at the EPA Region VIII Training Center in Denver, CO, November 9, 1989.

Mount, D.I. 1990. Number of test concentrations needed. *NETACommunique*. USEPA National Effluent Toxicity Assessment Center, Duluth.

Mount, D.I. and T.J. Norberg. 1984. A seven-day life-cycle cladoceran test. *Environmental Toxicology and Chemistry* 3:425-434.

Mount, D.I and T.J. Norberg-King. 1985. *Validity of effluent and ambient toxicity test for predicting biological impact, Scrippo Creek, Circeville, Ohio*. USEPA 600/3-85/044.

Mount, D.I., and T.J. Norberg-King. 1986. *Validity of effluent and ambient toxicity tests for predicting biological impact, Kanawha River, Charleston, West Virginia.* USEPA 600/3-86/006.

Mount, D.I., N.A. Thomas, T.J. Norberg, M.T. Barbour, T.H. Roush, and W.F. Brandes. 1984. *Effluent and ambient toxicity testing and instream community responses on the Ottawa River, Lima, Ohio.* USEPA 600/3-84/080.

Mount, D.I., A.E. Steen, and T.J. Norberg-King. 1985. *Validity of effluent and ambient toxicity tests for predicting biological impact on Five Mile Creek, Birmingham, Alabama.* USEPA 600/8-85/015.

Mount, D.I., A.E. Steen, and T.J. Norberg-King. 1986a. *Validity of ambient toxicity test for predicting biological impact, Ohio River, near Wheeling, West Virginia.* USEPA/600/3-85/071.

Mount, D.I., T.J. Norberg-King, and A.E. Steen. 1986b. *Validity of effluent and ambient toxicity tests for predicting biological impact, Naugatuck, River, Waterbury, Connecticut.* USEPA 600/8-86/001.

Mount, D.I., A.E. Steen and T.J. Norberg-King. 1986c. *Validity of effluent and ambient toxicity tests for predicting biological impact, Back River, Baltimore Harbor, Maryland.* USEPA/600/8-86/001.

Munkittrick, K.R. and E.A. Power. 1991. The relative sensitivity of Microtox®, daphnid, rainbow trout, and fathead minnow acute lethality test. *Environmental Toxicology and Water Quality* 6:35-62.

Murdoch, W.W. and E. McCauley. 1985. Three distinct types of dynamic behavior shown by a single planktonic system. *Nature* 316:628-630.

Nakamura, F. 1986. Avoidance behavior and swimming activity of fish to detect pH changes. *Bulletin of Environmental Contamination and Toxicology*. 37:808-815.

Norberg-King, T.J. 1988. *An interpolation estimate for chronic toxicity: The ICP approach.* Technical Report 05-88. National Effluent Toxicity Assessment Center, USEPA. Environmental Research Laboratory - Duluth.

Norberg-King, T.J. 1989a. An evaluation of the fathead minnow seven-day subchronic test for estimating chronic toxicity. *Environmental Toxicology and Chemistry* 8:1075-1089.

Norberg-King, T.J. 1989b. Review of the toxicity results from West Boise and Landers Street POTWs. Memorandum to Rob Pederson, USEPA Region X. USEPA Environmental Research Lab - Duluth.

Norberg-King, T.J. and D.I. Mount. 1985. A new fathead minnow (*Pimephales promelas*) subchronic toxicity test. *Environmental Toxicology and Chemistry* 4:711-718.

Norberg-King, T.J., and D.I. Mount. 1986. *Validity of effluent and ambient toxicity tests for predicting biological impact, Skeleton Creek, Enid, Oklahoma.* USEPA 66/8-86/002.

Pascoe, D., S.A. Evans, and J. Woodworth. 1986. Heavy metal toxicity to fish and the influence of water hardness. *Archives of Environmental Contamination and Toxicology*, 15:481-487.

Parkhurst, B.R., M.D. Marcus, and L.E. Noel. 1990. *Review of the results of EPA's complex effluent toxicity testing program.* Report to Utility Water Act Group.

Parkhurst, B.R. and D.I. Mount. 1991. Water-quality-based approach to toxics control: Narrowing the gap between science and regulation. *Water Environment & Technology* 3:45-47.

Parkhurst, B.R., W. Warren-Hicks, and L.E. Noel. 1992. Performance characteristics of effluent toxicity tests: Summarization and evaluation of data. *Environmental Toxicology and Chemistry* 11:771-791.

Patterson, P.W., K.L. Dickson, W.T. Waller, and J.H. Rodgers Jr. 1992. The effects of nine diet and water combinations on the culture health of *Ceriodaphnia dubia*. *Environmental Toxicology and Chemistry* 11:1023-1035.

Paulauskis, J.D., and R.W. Winner. 1988. Effects of water hardness and humic acid on zinc toxicity to *Daphnia magna* Straus. *Aquatic Toxicology* 12:273-290.

Pennak, R.W. 1989. *Fresh-water Invertebrates of the United States: Protozoa to Mollusca*. 3rd ed. John Wiley & Sons, Inc., New York.

Pickering, Q.H. 1988. Evaluation and comparison of two short-term fathead minnow tests for estimating chronic toxicity. *Water Research* 22:883-893.

Pifher, M.T., and J.T. Egan. 1992. WET Control: Square pegs do not fit in round holes. Presented to EPA's Water Quality Standards for the 21st Century seminar in Las Vegas, NV.

Plafkin, J.L., M.T. Barbour, R.D. Porter, S.K. Gross, and R.M. Hughes. 1989. *Rapid bioassessment protocols for use in streams and rivers* EPA-444/4-89-001.

Slobodkin, L.B. 1961. *Growth and Regulation of Animal Populations*. Holt, Reinhart and Winston, Inc., New York.

Snedecor, G.W., and W.G. Cochran. 1967. Statistical Methods. 6th Ed. Iowa State Unvi. Press, Ames.

Snyder, T.P., K.M. Switzer, and R.E. Keen. 1991. Allozymic variability in toxicity-testing strains of *Ceriodaphnia dubia* and in natural populations of *Ceriodaphnia*. *Environmental Toxicology and Chemistry* 10:1045-1049.

Sokal, R.R. and F.J. Rohlf. 1987. *Introduction to Biostatistics*. 2nd ed. W.H. Freeman and Company, New York.

Stewart, A.J., L.A. Kszos, B.C. Harvey, L.F. Wicker, G.J. Haynes and R. D. Bailey. 1990. Ambient toxicity dynamics: Assessments using *Ceriodaphnia dubia* and fathead minnow (*Pimephales promelas*) larvae in short-term tests. *Environmental Toxicology and Chemistry* 9:367-379.

Suter, G.W. II, A.R. Rosen, E. Linder, and D.F. Parkhurst. 1987. Endpoints for responses of fish to chronic toxicity exposures. *Environmental Toxicology and Chemistry* 6:793-809.

Thurston, R.V., R.C. Russo, and G.A. Vinogradov. 1981. Ammonia toxicity to fishes. Effect of pH on the toxicity of the Un-ionized ammonia species. *Environmental Science & Technology* 15:837-840.

Thurston, R.V., C. Chakoumakos, and R.C. Russo. 1981. Effect of fluctuating exposures on the acute toxicity of ammonia to rainbow trout (*Salmo gairdneri*) and cutthroat trout (*S. clarki*). *Water Research* 15:911-917.

USEPA. 1983. *Water Quality Standards Handbook.* Office of Water Regulations and Standards, Washington, D.C.

USEPA. 1991a. *Technical Support Document for Water Quality-based Toxics Control* (TSD) and Revision of 1985 edition. USEPA 505/2-90/001 Environmental Protection Agency, Washington, DC.

USEPA. 1991b. *Toxicity Identification Evaluation: Characterization of Chronically Toxic Effluent, Phase I (Draft).* EPA-600/6-91/005. Office of Research and Development, Duluth, MN.

USEPA 1991c. *Technicial Support Document for Water Quality-based Toxics Control-Responsiveness Summary.* EN-336. Environmental Protection Agency. Office of Water, Washington, DC.

USEPA. 1992. *Interim Guidance on Interpretation and Implementation of Aquatic Life Criteria for Metals.* Environmental Protection Agency, Washington, D.C.

Warren-Hicks, W. and B.R. Parkhurst. 1992. Performance characteristics of effluent toxicity test: Variability and its implications for regulatory policy. *Environmental Toxicology and Chemistry* 11:793-804.

Weber, C.I. (ed.) 1991. *Methods for Measuring the Acute Toxicity of Effluents and Receiving Waters to Freshwater and Marine Organisms.* 4th ed. EPA-600/4-90-027. USEPA Environmental Monitoring Systems Laboratory - Cincinnati.

Weber, C.I., W.B. Horning, D.J. Klemm, T.W. Neiheisel, P.A. Lewis, E.L. Robinson, J. Menkedick, and F. Kessler. 1989. *Short-Term Methods for Estimating the Chronic Toxicity of Effluents and Receiving Waters to Marine and Estuarine Organisms*, Second Edition. EPA/600/4-898/001. U.S. Environmental Monitoring Systems Laboratory, Cincinnati, Ohio.

Weisz, P.B. 1971. *The Science of Biology.* 4th Ed. McGraw Hill Book Co., New York.

Wilson, E.O., and W.H. Bossert. 1971. *A Primer of Population Biology.* Sinauer Assoc., Inc. Publishers, Sunderland, MA.

Winner, R.W. 1986. Interactive effects of water hardness and humic acid on the chronic toxicity of cadmium to *Daphnia pulex*. *Aquatic Toxicology* 8:281-293.

Winner, R.W. 1988. Evaluation of the relative sensitivities of 7-D *Daphnia magna* and *Ceriodaphnia dubia* toxicity tests for cadmium and sodium pentachlorophenate. *Environmental Toxicology and Chemistry* 7:153-159.

Woltering, D.M. 1984. The growth response in fish chronic and early life stage toxicity test: A critical review. *Aquatic Toxicology* 5:1-21.